Automated Theory Selection using Agent Based Models

Stratton, Robert James

Awarding institution:
King's College London

Automated Theory Selection using Agent Based Models

Robert Stratton

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London.

Department of Informatics
King's College London

February 24, 2015

Abstract

Models are used as a tool for theory induction and decision making in many contexts, including complex and dynamic commercial environments. New technological and social developments — such as the increasing availability of real-time transactional data and the rising use of online social networks — create a trend towards modelling process automation, and a demand for models that can help decision making in the context of social interaction in the target process. There is often no obvious specification for the form that a particular model should take, and some kind of selection procedure is necessary that can evaluate the properties of a model and its associated theoretical implications. Automated theory selection has already proven successful for identifying model specifications in equation based modelling (EBM), but there has been little progress in developing automatic approaches to agent based model (ABM) selection.

I analyse some of the automation methods currently used in EBM and consider what innovations would be required to create an automated ABM specification system. I then compare the effectiveness of simple automatically specified ABM and EBM approaches in selecting optimal strategies in a series of encounters between artificial corporations, mediated through a simulated market environment. I find that as the level of interaction increases, agent based models are more successful than equation based methods in identifying optimal decisions. I then propose a fuller framework for automated ABM model specification, based around an agent-centric theory representation which incorporates emergent features, a model-to-theory mapping protocol, a set of theory evaluation methods, a search procedure, and a simple recommendation system.

I evaluate the approach using empirical data collected at two different levels of aggregation. Using macro level data, I derive a theory that represents the dynamics of an online social networking site, in which the data generating process involves interaction between users, and derive management recommendations. Then, using micro level data, I develop a model using individual-level transaction data and making use of existing statistical techniques — hidden Markov and multinomial discrete choice models. I find that the results at both micro and macro level offer insights in terms of understanding the interrelationship between exogenous factors, agent behaviours, and emergent features. From a quantitative perspective, the automated ABM approach shows small but consistent improvements in fit to the target empirical data compared with EBM approaches.

Acknowledgements

I am grateful to many people for their support and collaboration in the course of my research. Firstly I thank my supervisors Peter McBurney and Michael Luck for their support and guidance. I would also like to thank the many people who have contributed feedback and discussion to my research, amongst them, in no particular order: Nazanin Jazayeri, Jinzhong Niu, Jennifer Castle, Ben Herd, Joanne Russell, Martin Chapman, Marnik Dekimpe, Dominique Hanssens, Vida Ghanei, Mehrdad Jazayeri, Laurence Tratt, Matthew Shaw, Padmaja Sasi, Laura Bennett, Samhar Mahmoud, Arash Eshghi, Gareth Tyson, Yassi Djazaeri, Jean-Pierre Muller, Simon Parsons, Sia Eshghi, Pierre Livet, Chris Sloane and Matthew Russell.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In many domains models are developed to represent a system or process with the aim of understanding and predicting dynamics in the target context. For example quantitative traders may create models of how the stock market works to try and maximise their returns, and meteorologists may develop models of climate change to understand and predict future temperature changes. These modelled representations can be used to help to support or refute existing theories and hypotheses, and may form the basis for decision making. The selection of the correct model for a particular domain may need to take into account a combination of empirical evidence and existing background knowledge about the field, and there may be a variety of possible models that could be used. Unless the true model is known, or a theory predefines a complete specification, some kind of selection procedure is needed in order to arrive at a specific model from the set of possible candidates. This can be a complex task, since there may be many candidates, trade-offs between prior theory and empirical evidence, and contending success criteria. Figure 1.1 illustrates this inter-relationship between theory, data and models — the diagram demonstrates a three way interaction, in which theory of a process is learnt through the validation of a representative model against the target data generating process.

In some domains, there are already partial solutions to the model selection task. In an Equation Based Modelling (EBM) context, involving models that are comprised of a set of variables and a linear or non-linear functional form which connects them [74], the computational issues resulting from testing the numerous models that arise from many combinations of variables, along with the desire to maintain certain characteristics in the selected candidates, have fuelled the development of automatic model selection processes [94]. These are said to outperform manual selection in terms of formulation, variable selection, unbiasedness of estimation, and their ability to consider all relevant evaluation criteria. Some of the automatic approaches used are simple stepwise routines, in which variables are sequentially added or deleted from an equation based on specific criteria [34]. Others, like PcGets and RETINA, are more developed approaches which aim to create models with particular statistical and theoretical properties [161].

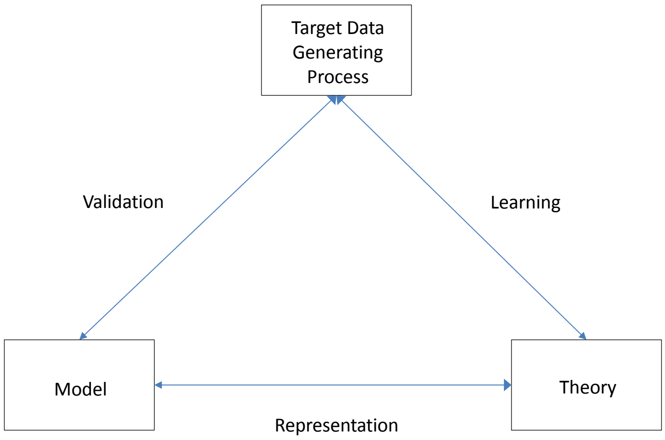

Figure 1.1: The relationship between theory, data and models

Reducing the size of the search space using existing theory makes the problem simpler. In common with other modelling methods, learning approaches to Agent Based Models (ABM — models in which the outputs are created from the bottom-up behaviour and interactions of individual simulated software agents [205]) can benefit from the use of an existing theory, or range of theories, in order to reduce the hypothesis space within which the parameters of the model are calibrated. Learning approaches in ABM can be broadly divided into those that focus on learning at a macro-level – in other words on the aggregated behaviour of the agents, and those that focus on micro-level learning. Macro level-studies, such as the Artificial Anasazi model [58, 102], aim to manipulate parameters to maximise fit against an expected outcome at an aggregate level, and are usually conducted against higher level aggregate patterns — often time series [149].

1.1.1 Elements Required to Solve the ABM Model and Theory Selection Problem

The main objective of this thesis is to develop and evaluate a solution to the problem of selecting an agent based model specification from the multitude of possible specifications that may exist. Although there are a number of broad approaches to model selection in the field of ABM, including the History Friendly and Werker-Brenner methods [205], no dominant methodology has so far emerged. In addition, there is a very diverse range of potential model selection tasks in ABM, depending on whether or not micro data is available, whether agents are assumed to be heterogeneous, and numerous other factors that may be relevant in specific applications. In this thesis a framework is proposed to enable automatic theory development and revision using agent based models as a tool to facilitate learning from data observed in the environment. Figure 1.2 shows the proposed components of this process which are explained in more detail below:

- **Automation Agent**— Automating the process of theory selection requires some kind of intelligent agent to administer the theory selection process and evaluate its outcome. Automation requires the system to take on, in a simplified fashion, a number of roles which in a manual modelling and optimisation process would normally be performed by humans. These tasks include the role of the thematician, who would normally manipulate theories, assumptions, and observations, and the role of model designer, who takes the thematician's conceptual model and creates a more formal design model [49]. In this case, the intelligent agent, acting as an automated analyst needs to be able to select a model of the processes at work in the environment from a set of candidate models, and interpret and take action based on the results of its model.

  The trend towards automation is driven by several forces including: a desire for speed — automated solutions may be able to work more efficiently within shorter product life-cycles [23] since manual approaches may have longer lag times; productivity gains — related in some ways to speed but also the ability to solve more complex problems; and superior decision making — due to improved quality and availability of data and the limitations of managers in information acquisition and processing [30].

- **Domain Theory**— Configuration for a particular domain enables existing theory to be brought to bear on the theory selection process. A theory is stated using terms from the domain's taxonomy and interconnects a set of laws into a unified theoretical account [50]. Domain configuration allows the researcher to introduce elements of their background knowledge into the system — knowledge about the domain that is separate from that specifically under study. The researcher usually considers this knowledge to be relatively certain, rather than the subject of active evaluation [50]. In addition the configuration phase allows the researcher to define the set of hypotheses that they wish to test to develop new knowledge in areas that they are less certain of. The user's theory might be based on, amongst other things, experience, empirical and theoretical literature, hunches, or results from qualitative studies [192]. In machine learning contexts, a factor which influences the definition or selection of the induced hypotheses is referred to as a bias [81] — by introducing a bias based on background knowledge of the domain, the user limits the hypothesis space.

  In practice it may not be possible for an agent to select from the total set of possible models — the number of candidate models may be extremely large, such that it is not feasible to compute all of the potential outcomes. A domain theory provides a structure with which to limit the hypothesis space that is to be tested. Parts of the theory may not be subject to testing as part of the modelling process. The theory may be a theory of individual or of collective behaviour, or a mixture of the two.

- **Theory-Model Mapping Rules**— As well as automating the roles mentioned above, clear commu-

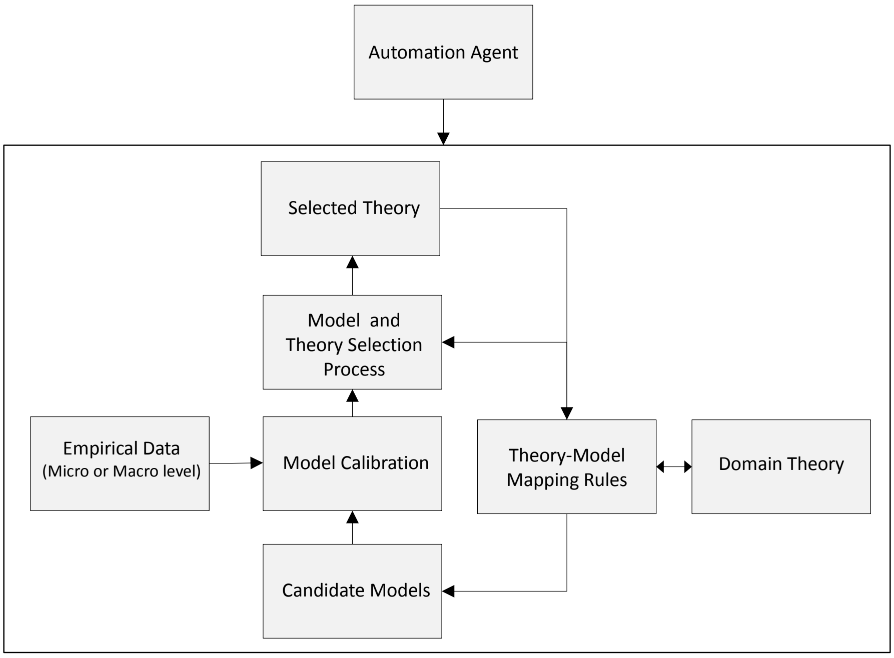

Figure 1.2: The elements of an automated model selection process

nication protocols between the different elements of the process also need to be defined and implemented. The mapping is a specification for how hypotheses in the theory specification can be articulated as a representation in a model. In addition it provides a framework for interpreting the outputs of a model in terms of the theory specification.

- **Candidate Models—** The term *model* is used in a large variety of different contexts, usually with a common theme that the model is a representation of another system or process. Some examples of models include scale models, which are smaller versions of the target, ideal-type models which exaggerate some features of the model and remove complicating factors and analogical models which are based on an analogy between a better understood phenomena and the target [74]. Models are developed and used with a variety of intentions including, amongst others, as illuminating abstractions, which can help the understanding of a process [59], as aids to develop explanatory theories of how causal processes work [62], and as a means to make predictions about how events in the real world may develop. In the current context the model is a computational representation of the framework defined in the theory specification, connecting the specification to empirical data in the target domain. There are often a number of candidate models because different theories and models might compete in order to explain a phenomenon of interest. The model framework needs to be able to implement the candidate hypotheses specified in the theory and to accept inputs and produce outputs which can be evaluated against empirical data. The model takes as inputs a set of empirically observed historic data, recorded over time at either individual or aggregate level. These represent the real world outcomes from a previous set of policy decisions – the real-world-data-generating process (*rwDGP*). The data includes independent variables that represent the exogenous factors that are hypothesised to play a part in the *rwDGP*, and a target variable which the model aims to approximate.

The form of the model that is used can vary. In many learning scenarios equation based models are used to represent the dynamics of an environment at an aggregate level. In other situations it may be the case that an ABM is the best approach to use. This is particularly likely where there is a need to model the individual actions of heterogeneous agents, where behaviours are expected to emerge from the interactions of agents, or a greater depth of theoretical representation is needed. Some authors, including Manzo [139], suggest that EBM demonstrates only the intensity and the sign of a relationship, while in contrast, agent based models emphasise *generative* mechanisms.

Figure 1.3 shows the relationship between the real, unknown, data generating process, the data that can be observed and collected about it, and an agent based representation of the process. The data includes a target behavioural variable which the model aims to approximate, and information about the environment at the different time points. The elements of the world that are believed to exist in the *rwDGP* correspond to elements of the agent model — which are represented through a set of core attributes to be parameterised. These attributes fall broadly into those that are related to the agents themselves, the links which make up the agent's interaction structures, and the environment in which the agents exist.

This kind of model can be used in at least four ways: to test an implication of a theory; to measure unknown values of theoretically defined variables; to predict the values of a variable; and to characterise a relationship and reveal relationships that become party of a theory [96]. The model may be built up from individual level behaviours at micro level, or collective behaviours at macro level.

- **Empirical Data**— Historic fit to empirical data, generated by the real data generating process, is one of the benchmarks against which the validity of the theory, mediated by the model, is assessed. Tsichritzis and Lochovsky define a datum as a triple (e, a, u) where the value u is selected from the domain of the attribute a to represent the attributes value for the entity e. Data is made up of a collection of these items [199]. A data representation can be defined as a set of rules for recording triples, and a data recording can be defined as a physical instance of a set of data items recorded using a data representation [65].

- **Model Calibration**— This involves searching the space of parameters of any given model specification using a search procedure such as grid search methods, gradient descent, simulated annealing or genetic algorithms. In this context, we define model calibration as the search, within the parameters provided by any given theory specification, to find the values for the parameters that collectively best explain the empirical data available. The values need to be determined collectively since correlation

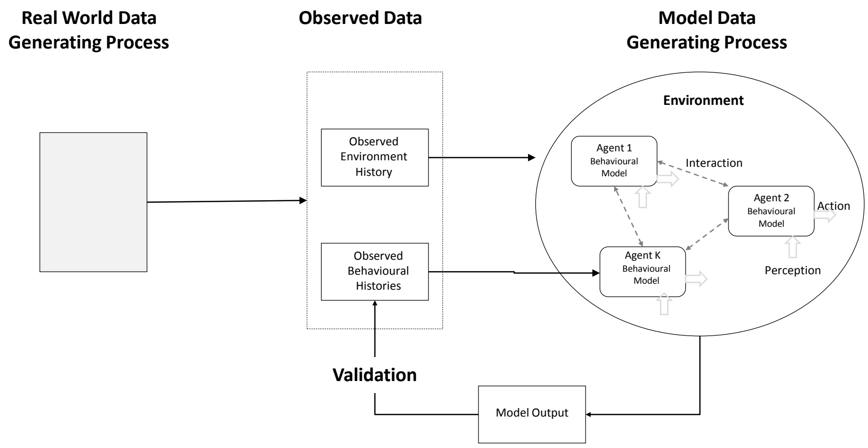

Figure 1.3: Parameterising an agent based model from observed data

between model inputs is a common and significant issue in model calibration, often resulting from a shared common time trend, a causal relationship between the inputs, or random variation [109]. When using correlated data, the components already included in a model can affect its current diagnostic criteria, as well as the sign and significance of the remaining candidate components [172].

- **Theory Selection Process** — This involves scoring the model against a number of criteria, including fit to the data —model accuracy, theory coherence, theory simplicity and congruence with existing beliefs. Two main criteria are considered for model selection: the qualities of the theory with which each model is associated, and the fit of the model output to the empirical evidence. As discussed above, the potential conflict between prior theory and empirical evidence can play an important role in model selection [94]

- **Selected Theory**— The selected theory is the one that performs best according to a given weighting of the different scoring factors.

The system is intended for use by a domain theorist, who supplies the theory specification, arranges the interface with the empirical data, and receives the selected theory. Figure 1.4 shows the interaction of the domain theorist with the system.

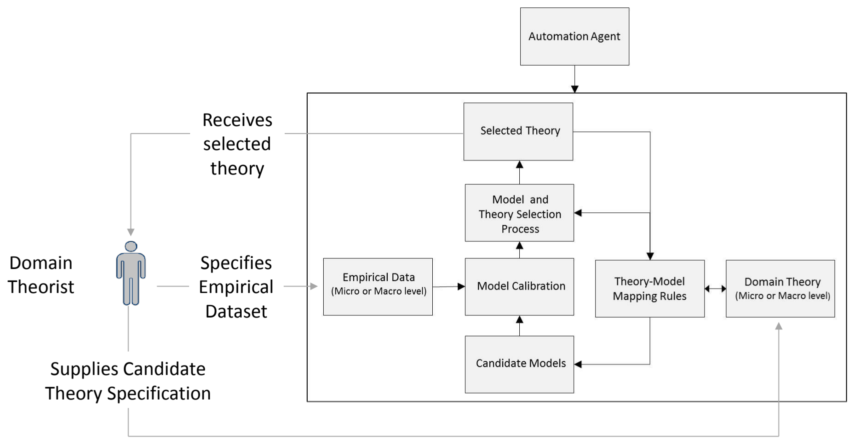

Figure 1.4: The relationship between the domain theorist, theory, data and models

1.2 Research Requirements

The primary objective of this thesis is to develop and evaluate a solution to the problem of selecting an agent based model specification from the many candidate specifications that may exist. The lack of a common methodological framework, and the potentially large number of models that may need to be specified when agents have heterogeneous characteristics, means that an automated model selection methodology could be at least as beneficial in an ABM context as it has been in EBM. The focus of this thesis is to develop an approach to automated agent based model selection that takes account of existing theoretical frameworks as well as empirical data at a micro or macro level. This overall objective entails various sub-objectives, listed below:

1. **Theory representation** — Automation requires a computational theory representation with which the modelling and data components can interact, and that is conformable with the specific characteristics of an agent based model, such as the capacity for events and behaviours to emerge at macro level from micro level behaviour

2. **Mapping theories to models** — As stated above, clear communication protocols between the theory representation and model need to be established that specify how a set of hypotheses expressed in the theory specification can be translated into a representation in a model.

3. **Model and theory scoring and criteria** — The candidate models and their associated theories need to be evaluated in a way that suitably balances theoretical conformity with the degree of fit to empir-

ical data.

4. **Model search mechanism** — The space of possible theories, models and their parameters resulting from the candidate theory specifications needs to be searched efficiently

5. **Interpreting a model in terms of a theory** — A framework needs to be created for the selected model and its associated theory to be automatically interpreted in a way that allows an appropriate action to be selected.

1.3 Methodological Approach

The success of the automated model selection method is evaluated through a combination of analysis of existing theoretical foundations for model selection, empirical evaluation against real world data sets, and evaluation of the theoretical interpretability of the resulting models. The approach is outlined below:

1. Establish a theoretical foundation based on existing work in the fields of philosophy of science, econometric model specification and machine learning

2. Develop system to reflect the theoretical foundations that can be implemented in a computational framework, working on real empirical data.

3. Execute the system over repeated simulations to evaluate its robustness

4. Evaluate the outcomes of these simulations against domain theory and empirical data

5. Compare the fit, explanatory power and theoretical congruence of the method against other rival methods

1.4 Research Contributions

To achieve the objectives listed above a system is developed that incorporates agent based models with a theory specification.

I claim to advance the state of the in the domain of theory selection using agent based models in the following ways:

- Adapting ideas from several fields into one coherent system for linking domain theory to agent based models, by defining an agent-centric theory representation which is able to represent agents' characteristics and their relationship with the corresponding exogenous and emergent factors which they perceive.

- By developing a set of criteria for scoring theories and models to produce an overall score by which a model and theory can be collectively assessed.

- By producing a customised representation of a genetic algorithm which is better suited to Agent Based Models through use of a non-binary alphabet.

- The method is verified against real data from two sources, one collected at macro level and the other at micro level, in which the data generating process involves interaction between users. Management recommendations are automatically derived that take into account the emergent characteristics of the environment.

- Issues relating to the role of social pressure, order dependence in social interaction modelling and memory in the attainment of equilibrium are explored using the modelling system and empirical data.

  The main contribution is conceptual, developing a framework which could be applied to a range of different domains.

1.5 Thesis Structure

Figure 1.5 shows an overview of the rest of this thesis. The thesis continues with a review of the literature in Chapter 2, then an evaluation of a simple implementation of an automated model selection system in Chapter 3. In chapters 4 and 5 a methodology is developed and described which is intended to provide a solution to the research question. The solution is then evaluated against different empirical datasets in chapters 6 and 7. The thesis is then concluded with a review of the evaluation and other methodological considerations.

The specific areas addressed in each chapter are detailed below:

- In Chapter 2, the existing literature is reviewed in the areas of intelligent agents, theory, modelling approaches in general, specific features and agent based modelling. In addition, some of the issues that are relevant to the particular domain in which the research is applied — marketing decision making are introduced, and describe the current modelling methods used to support these decisions. I review existing work in automation and social interaction representation for EBMs, and consider how these could inform the development of an automated ABM-based theory selection process, suggesting why they would be appropriate and considering how they would need to differ from analogous components in an EBM methodology.

- In Chapter 3, The performance of automated model selection processes using ABM and EBM is compared in a series of simulated encounters between manufacturers in a simulated market, with the rival manufacturers using contending approaches to evaluate their decisions.

- In Chapter 4, approaches are proposed to two of the research requirements raised in Section 1.2. I describe an agent-centric theory representation that would form the basis for an automated modelling framework using ABM, which takes into account the emergent characteristics resulting from agent

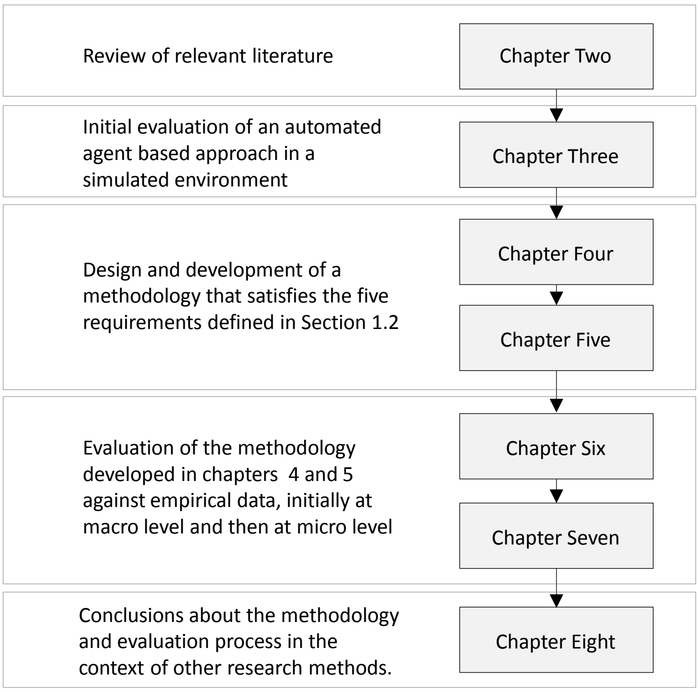

Figure 1.5: Overview of the thesis structure

interactions. I then detail a rule-based mapping protocol that would allow the theory specification to be mapped to different types of ABM to allow testing across a variety of methods and applications.

- In Chapter 5, solutions are proposed to the additional three research requirements identified in Section 1.2. I propose a set of scoring methods that could be used to evaluate the qualities of a theory based on the representation described in Chapter 4. I then identify an approach to searching the theory and parameter spaces involved in a particular context using a genetic algorithm. Finally I describe a rule-based method for action selection based on the theory selected.

- In Chapter 6, the overall framework developed in Chapters 4 and 5 is evaluated at a macro level. The evaluation uses real data from an online social networking site in which the data generating process involves interaction between users, and derives management recommendations that can be used to optimise future performance.

- In Chapter 7, the framework is extended to work against data collected at micro-level, mapping the theory specification to existing statistical methods and using observed behavioural data to develop models of individual agents and their interactions.

- In Chapter 8, the benefits of the overall framework compared to existing methods are reviewed, and the outputs compared with other work in related areas. I also suggest some areas that could be further developed in future work.

Chapter 2

Literature Review

2.1 Introduction

This chapter provides an overview of recent literature in the areas that are relevant to this research — the development and evaluation of a solution to the problem of selecting an agent based model specification from the numerous potential specifications that may exist. The full set of subsidiary research requirements entailed by this were laid out in Section 1.2.

Since the literature that is relevant to these requirements spans a number of sub-disciplines, this chapter is broken up into several sections. In Section 2.2 some of the core characteristics of agents and agent based models are reviewed. Then in Section 2.3 recent work on agent based models is considered, covering the main areas that are relevant to this research — especially interaction and complexity. In addition, ABM is situated in the context of other types of equation-based and computational models. These areas are relevant to the automation of the process and the definition of the candidate models that the process is choosing between. In Section 2.4 the literature on selection, validation and verification in ABM is reviewed, with particular reference to the statistical features of the models. This section is particularly relevant to the model calibration task and the mechanics of model and theory selection. Section 2.5 examines some of the issues relating to theory in the context of ABM, and broader considerations for choosing between theories in scientific research. This is particularly relevant to the representation of a domain theory, and also theory selection. Finally, Section 2.6 gives an overview of marketing concepts and recent applications of ABM in marketing contexts. Collectively these sections provide a review of the literature in fields relevant to automated model specification in ABM.

2.2 Agents

In this section some definitions of agents and their characteristics are discussed, and a number of approaches to agent decision making are considered. General agent concepts are relevant to the research objective since agents form the basis of agent based models — which are at the core of the modelling process used in this research — and are also a means for implementing the automated model search procedure.

2.2.1 Agent Definition

There is currently no universally accepted definition of an agent. Different domains and applications have tended to pursue separate objectives and architectures, leaving relatively few areas of overlap. For example, from an AI perspective, Russell and Norvig describe an agent as ‘anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators’ [172]. From a multiagent systems viewpoint, Wooldridge suggests a requirement for autonomy, and also the notion of an objective delegated by some other party [206]. For the purposes of this work, the definition put forward by Wooldridge and Jennings [207] will be used. They identify a set of properties that can be considered to characterise an agent, suggesting that an entity can be considered to be an agent if it has the following:

- **Autonomy** — the capacity to operate without the intervention of humans or others, and to have some control over its actions.

- **Social Ability** — the ability to interact with other agents, and potentially also humans, using some language.

- **Reactivity** — the capability to perceive its environment and react reflexively to changes in it.

- **Pro-activeness** — the ability to take the initiative and exhibit goal directed behaviour.

Other properties may be more important in certain applications, for example mobility — the capacity to move around a network, or the ability to store memories.

2.2.2 Agent Behavioural models

Having reviewed some definitions of agents and their characteristics, this section looks more specifically at their behavioural and decision making capabilities. Three specific approaches to agent behaviour are considered — PRS, subsumption and SOAR.

2.2.2.1 PRS

The Procedural Reasoning System was developed at Stanford Research Institute during the 1980s by Georgeff, Lansky and others. Now one of the most established BDI implementations, PRS has been used in a variety of real world applications including a role as a monitoring and fault detection system for the NASA space shuttle, air traffic control systems and others [206]. The beliefs desires, and intentions of an agent are not explicitly represented in PRS, but the elements of BDI can be ascribed to it: the agent’s beliefs are considered to be the information that it has about the world, which is not necessarily correct; the agents desires are its goals, its intentions are the subset of desires that it will commit resources to [206, 207].

The PRS agent observes the world and is sensitive to changing events. An event can be a new belief or a new goal. The agent maintains a set of events which it matches with a knowledge database of plans that specifies how it may react to events or bring about a state of affairs. The plans are pre-written by the programmer at design time. The events that it has matched with possible plans are considered to be its desires. Plans in PRS have a goal (the postcondition), a context (the precondition) and a body, which is the course of action to carry out. The body of a plan is split into goals, subgoals and primitive actions. The agent selects one of the plans for execution which is considered its intention. Once a goal is achieved or it becomes obvious that it can not be achieved, a new goal is selected [167, 206].

There is a c++ implementation of PRS, known as the distributed Multi-Agent Reasoning System (dMARS), and a simpler, textual language, *AgentSpeak*, that uses a first-order syntax with events and actions [167]. PRS limits computational complexity through the fact that agents do not do any first principles planning, and its recognition that since in general an agent will not be able to achieve all its plans, it will choose a subset of plans to execute. Agents have no capacity to learn in PRS since their library of plans is prebuilt [206].

2.2.2.2 Subsumption

The subsumption approach was developed by Rodney Brooks, a researcher at MIT. He outlined his approach in a 1985 paper *A robust layered control system for a mobile robot*. He was critical of symbolic AI's reliance on the assumption that a world model could be built internally and then manipulated [28]. Brooks felt that existing approaches were inherently unrealistic, arguing that real biological systems are not rational agents that take inputs, compute logically, and produce outputs. He proposed that actually, intelligent behaviour does not require explicit symbolic representations, or explicit abstract reasoning. Instead the subsumption approach uses a layered set of behaviours, analogous to a biological nervous system, in which higher level layers seek to subsume the lower levels.

There are several concepts that underlie his research. Brooks suggests that real intelligence is situated in the world, and that intelligent behaviour emerges from an agent's interaction with its environment. He argued that intelligence is emergent from certain complex systems, emerging from the bottom-up rather than from top-down central control [207]. Lower behavioural layers are more primitive, designed, for example, to react to the environment to avoid obstacles. Upper layers are more abstract and contain broader objectives. Each layer is represented by an Augmented Finite State Machine — the machines are augmented by their use of timers. The behaviours compete to gain control of the robot. The behaviours act as reflex systems — given a particular input they will produce a particular output, and there is no adaptable response. No reasoning process, nor inner states or preferences modify the response to a particular input. Unlike some architectures, there is no central controller, behaviour is emergent rather than governed [206].

Brooks built a number of convincing robots using his approach, and in some scenarios achieved aggregate behaviours that rivalled more complex architectures. Advocates of the approach argue that subsumption is better able to cope with imperfect sensory data or an unpredictable environment. The multiple behavioural layers that it uses may tackle a problem separately so it is more likely to fail in stages than in a single event. Also because agents conduct no explicit reasoning, the computational demands of the architecture are relatively low [206].

Criticisms of subsumption include the fact that because it is hard for a designer to understand all of the possible interactions between layers, unexpected or undesired behaviours may evolve. Despite the apparent desirability of emergent behaviour, work may be required to make sure that certain behaviours do not appear. Others suggest that because of the absence of a central controller, it is harder to change the agent's task.

2.2.2.3 SOAR

Soar (State, operator and Result) has been under continuous development since the early 1980s and has been used in application since 1983. Most of the development has been carried out at the University of Michigan. Some of the applications it has been used for include modelling the actions of fighter pilots in aerial combat scenarios, guiding intelligent agents in computer games and modelling categorisation problems. Soar does not seek to replicate the workings of the brain or its anatomy [121]. All tasks in Soar are an attempt to achieve a goal, with goals represented in working memory. The tasks are characterised in terms of the selection and application of operators to a state.

Operators may refer to simple or abstract tasks, and the selection of which operator to apply depends on the application of knowledge, memory and preference. Production rules are matched with working memory to bring relevant knowledge into play. A state in Soar refers to the current problem-solving situation. The decision cycle selects appropriate actions, bringing the system closer to its goal. If a goal can not be directly achieved it is broken down into sub-goal [120, 38]. Impasse occurs in the decision cycle when knowledge about operator selection is not sufficient. Soar learns from impasse through a combination of:

- Chunking occurs when one or more results are produced in a sub-goal. Soar learns a chunk as a production rule that can directly create the results of the processes involved. This rule can be applied in similar situations, rather than the sequence of rules.
- Reinforcement learning happens through rewarding an operator's score for successes and penalising them for failure. Operators that have received a reward are more likely to be used in the future.
- Episodic and Semantic memories store past experiences. Because they are used to select relevant operators, the dynamically changing history they reflect will modify the selection outcome.

Soar uses separate memories and representations for descriptions of its current situation and its long-term knowledge. Long term memory includes procedural memory which contains information about how to complete tasks, semantic memory which contains previously known facts and episodic memory which holds a history of previous states. Working memory contains the current situation, including data from sensors, results of intermediate inferences, active goals, and active operators [121].

2.2.2.4 ACT-R

ACT-R has been in development since the 1970s, and has gone through a range of enhancements. It is relatively unique amongst mainstream cognitive architectures in that it aims to create a model of human cognition that can recreate the outcomes of cognitive experiments on humans [121]. Also, since its modules and production system are mapped to actual brain regions, it can be compared against actual brain-imaging information to predict brain activation patterns in certain conditions. ACT-R has been applied to various tasks including the identification of enemies in combat simulation and air traffic control [38].

Long term memory contains declarative knowledge in the declarative module and procedural knowledge in an intentional module. The modules in the system are coordinated by these production rules, with each production rule having an associated cost and probability of success. The contents of the short-term memory determine which rules will be fired, taking into account the expected costs and benefits of each. The different modules only have access to the information held in the various buffers not the full set held in the module itself. For example, perception is stored in the visual module and partially made available through visual buffer and goals stored are stored in the intentional module and made available through the goal buffer. Working memory contains goal, perception, relevant knowledge, and motor action in the various buffers [38, 121]. Learning in ACT-R occurs through [6]: the rewarding of frequently used declarative chunks over less used ones; the updating of the cost and benefit of production rules based on their actual performance; and the combining of multiple input conditions and rule firings into new combined rules.

2.3 Agent Based Models

In this section agent based models are defined and discussed along with a comparison with other types of equation-based and computational models. The literature concerning some of the major features of agent based models is then reviewed, covering the role of the environment and interaction between agents. Finally, important areas of recent research in agent based models are then reviewed, covering complexity, trust, cooperation, reputation and norms. Agent based models are central to the research since they are used to represent the theory that is being investigated, with their outputs validated against empirical data. The form and structure of the model is important since it needs to be able to reflect a theory, and needs to be parameterisable by the search process.

2.3.1 Model Definition

Models are a central feature of this research. Livet [133] cites Minsky's [145] definition of a model: 'To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A'. For Gilbert [74], the real system that is represented is the target of the model. He defines four types of model: scale models, which are a smaller version of the target and are associated with a likely loss of detail; ideal-type models which exaggerate some features of the model and remove complicating factors; analogical models which are based on an analogy between a better understood phenomena and the target; and equation-based models which specify relationships between variables such as structural equation models.

2.3.2 Other Types of Model

2.3.2.1 Equation Based Modelling

Equation based models are referenced in later chapters of this thesis, primarily as a comparison with the performance of agent based models. Equation based models are comprised of variables, typically selected to be representative of the target phenomenon, and a functional form which represents the relationship between them. The equation might take a number of forms, both linear and nonlinear. Gilbert [74] cites the example of the Cobb Douglas production function:

$$Y = AL^{\alpha}K^{\beta}$$

where $Y =$output, $L =$ labour input, $K =$capital input and A, α and β are constants that represent technology.

The equation above reflects an economic theory that the relationship between the dependent variable output, and the inputs labour and and capital, is multiplicative rather than additive, reflected in the form of the equation $Y = AL^{\alpha}K^{\beta}$.

Gilbert [74] argues that the form of the equation is of little consequence in such models, and that the emphasis instead is on making the outcome fit the data. Manzo [139] suggests that statistical models demonstrate only the intensity and the sign of a relationship. Both conclude that these types of model are not helpful in determining the underlying process and mechanism behind the process. In contrast, agent based models are considered to emphasise generative mechanisms. Epstein [58] argues that there is no fundamental distinction between agent based models and equation based models, since any agent based model can be reduced to a set of recursive functions. He argues that although these are currently difficult to understand or interpret – there is a possibility that their use will become widespread. Epstein [57] also argues that in many cases the same result may be available using an equation based model, but that the equation based model offers a descriptive rather than generative approach. Macy [137] concurs that ABM is more concerned with the mechanisms that underlie the process than predictive accuracy.

Parunak *et al.* [159] see one of the fundamental differences between equation based and agent based models as being the level of aggregation at which the relationship between the elements operates. Equation based models use aggregate system-level variables, while in an agent model these aggregations emerge through lower-level behaviour and may not be directly accessible to the agent. Axtell [12] points out that there is an implicit assumption in equation based models that social agents interact not with each other, but with abstract economic objects like price vectors and unemployment rates. Many authors agree that one of the key comparative strengths of the agent based approach is its explicit handling of these agent interactions [74, 12, 58, 159]. In addition it is possible to represent heterogenous actors, and impose modes of behaviour such as bounded rationality on the actors [12]. Macy [137] argues that in some simulations it makes sense to incorporate aggregate variables that are observable to the agents, such as climate. Frigg and Reiss argue that computational models do not relate to their target systems in a way that is significantly different to other modelling methods. They suggest that they create no new philosophical problems beyond those already presented by activities such as thought experiments and mathematical models [68].

2.3.2.2 Simulation Based Models — Overview

Manzo [139] defines simulation as the execution of a program that translates a theoretical system (representing an object of analysis) into a set of algorithms written in a specialized computer language. Simulation is a broad family of methods that has been put to a wide range of uses. From an epistemological viewpoint, simulations introduce an intermediary step between a theory and a phenomenon [132]. Gilbert and Troitzsch [76] identify an array of objectives for simulation, including understanding, prediction, substitution for humans, training and entertainment. A key feature of the development of simulation is that it has closely followed the development of computing. As hardware and software have improved, so too have the capabilities of computer simulation. Macy [137] identifies three periods of development in social simulation. He argues that the 1960s saw the development of macrosimulation, concerned with flows, control and feedback processes amongst higher level-entities like warehouses and countries. He suggests that in the 1970s, there was more use of microsimulation, using individual level units of analysis to simulate the movement through time of specific conditions. Agent based modelling appeared in the 1980s, and was distinct in that it allowed interaction and interdependence between agents.

2.3.2.3 Monte Carlo Simulation

Monte Carlo Simulation generates artificial data in a way that aims to reflect the theoretical data generating process in the real world, and so help statisticians to understand how probability sampling distributions impact on a model's results. It involves creating a pseudo-population, with each member of the population having random and deterministic characteristics. The random characteristics are typically drawn from a specific distribution. With each run of the simulation, agents in the pseudo-population generate outputs from a given random distribution, such that the overall distribution reflects that of the true population being modelled [147]. Axtell [12] argues that in Agent Based Models which use randomness, there is a distribution of possible outcomes which is analogous to the outcomes of a Monte Carlo simulation.

2.3.2.4 System Dynamics

Lattila *et al.* suggest that the Systems Dynamics approach was developed by Forrester in the late 1950s [122]. It is used to model systems of interacting variables and to characterise the cause and effect relationship between these. The future state of the system is derived from its current state. One of their strengths is that they can be used to represent feedback loops. But unlike agent based models, systems dynamics deals with sets of group aggregates e.g. wolves and sheep in an eco-system model, rather than the individual entities — sheep and wolves. Because of its lack of ability to directly represent individual agents it can not support diverse behaviours at sub-group level diversity can only be built in by introducing

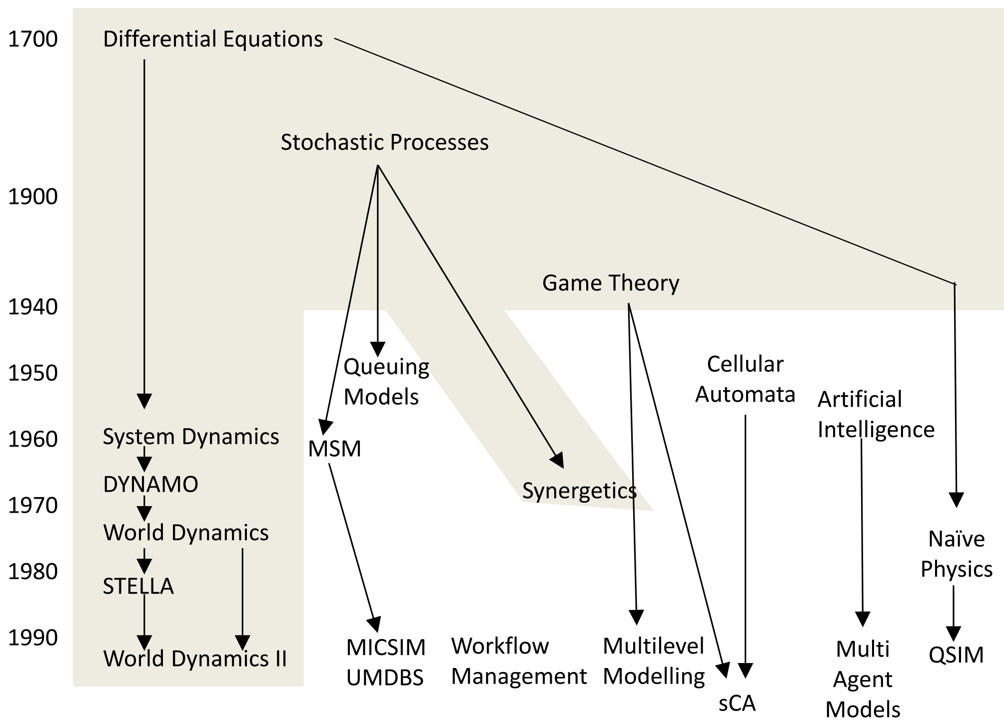

Figure 2.1: The history of simulation based methods — from [76]

additional groups [74]. Lattila *et al.* [122] argue that although the approach does not deal with complexity in the same sense as agent based modelling, it embodies a dynamic complexity which develops over time due to the nonlinear feedback loops which exist within the modelled systems.

2.3.2.5 Discrete Event Simulation

Gilbert and Troitzsch [76] suggest that in many disciplines, like engineering and workflow management, discrete event modelling is the dominant form of simulation. For Garrido [72], a discrete event model is one that changes its state at discrete points in time. He cites the example of a customer arriving at a barber shop as an event at a specific instant in time. Time does not pass at as equidistant time-steps, but is determined by events between events, nothing happens which is of relevance to the model [76]. The event changes the state of the model in terms of the number of customers waiting for a haircut. Garrido contrasts this approach to time with that of a continuous model in which changes in state occur continuously for example the temperature of a boiler in a powerplant. Gilbert and Troitzsch [76] argue that discrete event models are most widely used to represent queuing scenarios, with the key objects being servers, customers and queues. The times between arrivals of customers are generally stochastic.

2.3.2.6 Microsimulation

Macy defines microsimulation as an approach which models the development of simulated entities over time [137]. Applications of microsimulation are focussed in the area of social policy, for example the artificial ageing of a set of households to evaluate how its income and pension requirements may change in the future. The starting point for the simulation is often a set of real conditions based on data collected through a survey. Unlike agent based models microsimulation does not accomodate dynamic adaptation or interaction between entities. Amongst the challenges of the method is that it requires a detailed application of probabilities at each time increment, for example the probability that a household member will become unemployed in a particular year, and these probabilities may not be independent but conditional on multiple conditions e.g. age and previous work history [74]. More sophisticated microsimulations may include multiple levels of aggregation for example individuals may be grouped into households as well as having individual characteristics [76].

2.3.2.7 Cellular Automata

First proposed by John von Neumann in the 1940s, Cellular Automata are used to model interactions between cells [192]. The cells are arranged in two or three dimensional grids, and each cell can be in one of a number of different states, for example *dead* or *alive*. Time proceeds in steps, with rules determining the state of a cell in any given time period [76]. The rules are usually simple and interactions local — within the immediate neighbourhood of the cell [192]. Cellular automata are used to investigate the outcomes of simple micro-scale events and have been used in physics, biology and other disciplines. A well-known example of a cellular automata application is Conway's Game of Life. Each cell in the model uses two simple rules it dies unless it has two or three living neighbours, and it comes to life if it has exactly three living neighbours. The simple micro rules lead to emergent patterns at a macro-level [76].

2.3.3 Objectives for using Agent Based Models

Livet [132] proposes a variety of conditions data and theory conditions that a modelling enterprise might begin from. These range from a situation in which the modeller has a formal theory but can not say what theory would predict by analytical means because of number of variables or because of interactions, through to the situation in which he has no theory of the domain only some of the phenomena and typical scenarios observed in the data. The following categories incorporate a wide range of circumstances in which agent based models are used:

- **Improving empirical understanding and prediction** The investigation of whether large-scale regularities have emerged from low-level processes and how reliably they can be predicted [11].

- **Normative understanding** The study of how models can be used to improve the design of policies in different settings [11]. As part of this process a tool might be created to support decision making amongst possible futures [62].

- **Heuristic** This is concerned with improving understanding about the underlying causal mechanisms at play [11]. Epstein argues that models can be helpful as illuminating abstractions in so far as they capture qualitative behaviours, without necessarily being completely accurate [59]. He argues that there is an important distinction between explaining and predicting, with explanation having a value without a necessity for it to predict. He cites as an example the theory of plate tectonics. The theory explains why earthquakes happen but does not necessarily predict their time and place. As part of this exploration of causal mechanisms, Ferrand [62] mentions the ability to test hypotheses about system dynamics. As a sub-goal of explanation, he includes education in the sense of providing a platform for training.

- **Methodological advancement** Methodological advancement encompasses a range of sub-goals including: exploring how tools and approaches can be improved, enforcing a scientific habit of mind and guiding data collection [59]. Epstein argues that there is a cyclical process through which the process of observation and theoretical development informs the collection of data to support these observations and theories.

- **Facilitating interaction and cooperation** Ferrand identifies a number of situations in which agent based models can be used with multiple stake holders. As an example he suggests that in an industrial setting like a supply chain, a modelling exercise may be undertaken between multiple parties with each party maintaining the confidentiality of their own processes by hosting their own agent. In other situations he suggests that communication between stakeholders can be improved through the process of jointly building a model [62].

Various studies have reviewed the strengths and weaknesses of different types of interaction model, primarily from the perspective of innovation diffusion. These studies point to a number of advantages in using ABM over equation based methods. As discussed in Section 2.3.5, in agent based applications, interaction can include a wide range of possible types of exchange between agents such as collaboration, negotiation [165] transfer of data, and the communication of messages [74]. The interaction structure defines the links between agents in the model, and could be taken from an existing network structure or a variation of it. Axtell [12] points out that there is an implicit assumption in equation based models that social agents interact not with each other, but with abstract economic objects like price vectors and unemployment rates. Many authors agree that ABM is better able to take account of differing consumer characteristics and different structures of social interaction [111, 74, 12, 58, 159]. Aggregate equation based models have been criticised for their failure to explicitly consider consumers' heterogeneity, for ignoring the complex dynamics of social processes that shape the diffusion, a lack of predictive and explanatory power [111] and for the over-simplification of the external influence parameter [100]. Goldthorpe suggests that the ability to recreate a series using a generative process — a process acting at a more microscopic level which is creating the observed data — is a useful addition to other methods. He suggests that equation based statistical modelling can show the impact of smoking on lung cancer, but in proving a causal link the identification of carcinogens and their physiological impact was crucial [80].

2.3.4 Environment

According to the basic definition of an agent cited earlier, the environment is a necessary precondition for several core behaviours, including perception [172] and reactivity [207]. Gilbert [74] defines the environment as the virtual world in which the agents act, which may have a greater or lesser impact on their behaviour depending on the application. For Odell *et al.* [157], the environment provides the conditions under which an agent can exist. Weyns *et al.* [203] suggest that the environment in an agent based model is generally considered to cover three broad types of structure: physical, communication and social.

For relevant simulations, the physical environment can include physical laws such as gravity, as well as temperature, humidity and any other items that are pertinent. Bandini *et al.* [17] give an example of a model in which agents are particles which are subject to and also generate forces. In this case, they argue, the environment determines the overall dynamics of the system. The physical environment may also impose constraints concerning which areas are accessible to an agent, by what means the agent is able to move around, and the degree to which an agent can modify its surroundings [157]. Weyns *et al.* [203] point out that the environment is also important for agent interactions, it can either enable or constrain them.

Gilbert [74] refers to environments which reflect a geographical space as spacially explicit. In these models, the spatial distribution of agents in an environment can affect the probability that agents or groups of agents will meet or neighbour each other. In some cases, like Schelling's simulation of residential segregation, spatial proximity is central to the development of the model's dynamics. In other models the environment may not be physical but instead reflect a knowledge space, in which proximity represents a similar attitude or viewpoint.

Weyns *et al.* [203] argue that as well as playing a structuring role, the environment is a first class abstraction – an independent building block that has its own responsibilities. They argue that there should be a clear separation between the roles of agents and the environment, with the environment used to embed resources and services, and to independently maintain its own dynamics. They also suggest that there are two distinct perspectives on the environment one for situated agents and another for cognitive agents. Situated agent systems have more of a history of exploiting the environment. They may also use the environment as a coordination medium and to share information. Stigmergic coordination is the manipulation of marks in the environment, for example digital ants leaving digital pheromones that form paths to locations.

2.3.5 Interaction

Interaction is a key feature of the simulations presented in later chapters of this thesis. Interaction is widely considered to be at the core of multiagent systems, and incorporates a wide range of possible types of exchange between agents. These include collaboration, negotiation [165] transfer of data, and the communication of messages [74]. In addition, interaction could refer to events in which agents respond to the presence or proximity of other agents without an exchange. As a central element of a multiagent system, interaction incorporates or is closely related to other areas of multiagent theory. Interaction also forms the basis of complex systems and their emergent features. In some applications, interaction is affected by an agent's disposition towards other agents, including their willingness to trust. Disposition may in turn be influenced by discernible agent characteristics such as tags [168], a history of experience with the same agent [10] or information about the agent obtained from other agents or the environment [156]. As well as determining the likelihood of interaction who an agent chooses to interact with [168], and what actions to take if an interaction does occur [113].

The probability of interacting with any particular agent is also affected by the network topology, for networked agents, or the environment. For example physical barriers may make it impossible for particular sets of agents to meet [203]. In a population of heterogeneous agents, successful interaction relies on mutually understood protocols, languages and ontologies [165].

2.3.6 Communication

In many multiagent systems, agents need to exchange knowledge, therefore the ability to effectively communicate is an important property. Well known applications of agent communication include e-commerce components that request proposals from automated suppliers. In order to communicate meaningfully, agents need a common specification for how messages will be sent and received, as well as their interpretation. Singh (1998) proposes three key aspects of language that should be considered in any formal analysis:

- **Syntax** — relates to the symbols contained in a communication and how they are structured.

- **Semantics** – concerned with what the symbols denote.
- **Pragmatics** — how the semantics are interpreted and used.

Singh suggests that meaning arises from a combination of semantics and pragmatics [186]. In many approaches to agent communication, speech is treated as an action agents perform speech acts, in the same way as they would other actions, to further their intentions. Wooldridge cites John Austin's (1990) assertion that speech is an action in the sense that it changes the world e.g. the sentence 'I declare you man and wife', spoken by the appropriate person, changes the real state of the world. Austin identified various performative verbs like request, inform, promise [206].

Like other actions, speech actions may fail, because successful communication can not be guaranteed by the sender alone, the receiver has some control over their reaction and whether the objective is achieved. When considered as an action, communication can be built into planning processes as part of a series of tasks intended to achieve some goal.

Many applications use a proprietary Agent Communication Language (ACL), meaning that although agents running on the same architecture can communicate, they are not interoperable with others. There have been several attempts to create a common ACL. Singh [186] identifies a core set of statement categories that will be a useful basis for an ACL in computing scenarios:

- *Assertives*, which inform:
- *Directives*, which request:
- *Commissives*, which promise something
- *Permissives*, which give permission for an act
- *Prohibitives*, which ban some act
- *Declaratives*, which cause events in themselves
- *Expressives*, which express emotions and evaluations

In the late 1980s the US agency DARPA helped to develop one of the most successful ACLs — the Knowledge Query Management Language (KQML). KQML assumes that each agent has a virtual knowledge base, and provides a syntax that allows assertive and directive communication. Criticism of the KQML approach focused on the number of varying dialects that developed, and the difficulty of ensuring that applications were semantically as well as syntactically compliant [206].

KQML has largely been superseded by the FIPA (Foundation for Intelligent Physical Agents) standard, which was based on the Arcol specification created by France Telecom in the 1980s. Arcol and FIPA have fewer primitives and enforce a semantic definition based on an agents beliefs and intentions [206]. Singh suggests a number of problems that can arise from communication in FIPA where these conditions are ambiguous [186].

2.3.7 Social Networks

In domains like opinion dynamics and peer to peer file sharing, networks are key to defining which agents may interact and what interactions are possible. Representations of social networks usually contain at least two key components: actors and relations. An agent might be any relevant entity, such as a firm, a country, or an individual. The relation may represent one of a number of possible relationships that can exist between agents. Knoke and Yang outline a typology of possible relationships that includes Communication relations through which messages are exchanged, and Transaction relations in which agents exchange control of physical or virtual entities. There may be multiple types of relationship between a pair of agents and each agent might be part of multiple networks [116].

Formal definitions of networks draw on concepts from graph theory, using the term node to refer to agents, and edges to refer to relationships. Edges with an arrow facing in one direction are directed representing a one way relationship. Edges with no arrows, or an arrow at each end are used to represent undirected, mutual relationships. Different types of line and arrows may be used to distinguish the various possible relationships [116]. For a node, the number of head endpoints adjacent to a node is called the indegree of the node and the number of tail endpoints is its outdegree.

2.3.7.1 Network Topology

The overall structure created by the combination of various edges and nodes forms the topology of the network. Topologies can be completely random, with every node as likely to connect to any other node, or regular, marked by dense groups of local connections. Useful summary information for a network are average path length, where the path length refers to the lowest number of edges connecting two nodes and a clustering coefficient, which describes the percentage of possible edges between neighbouring nodes that exist.

Watts and Strogatz describe how they developed a hybrid network by randomly rewiring some of the connections in a regular, clustered network. They found that by introducing a relatively small number of random connections between clusters they could recreate similar path length characteristics to that of a random network [202]. This type of network is best known as a small world, and can be demonstrated using the patterns of co-starring occurrences from the IMDB film actors database. The Kevin Bacon game is often used to illustrate how Small Worlds work. In the game, players have to find the connection between a particular actor and Kevin Bacon by linking their co-starring film roles [202].

2.3.7.2 Game Theory

Game theory is relevant to a review of interaction in that is a widely used approach to formalising the possible gains and losses that each agent may achieve during an encounter with one or more other agents [206]. Such interactions are usually assumed to be between self-interested agents who have different levels of utility for the possible outcomes from a situation. The outcome may be affected by the actions of the agent, and also the actions of the other player. Utility can be divided into three sub-sets.

- **Ordinal utility** is just a comparative preference for one thing over another, for example driving over walking, but without the strength of the preference being defined.
- **Cardinal utility** defines the strength of preference by applying a score to each thing, for example driving scores 10, walking scores 6.
- **Expected utility** expresses the average utility that could be expected given the probability of a confounding event. Hargreaves and Varoufakis give the example of an individual who has a higher utility for walking over driving when there is no rain, but in the event of rain would much prefer to drive. Their expected utility for walking would take into account the perceived probability that it would rain. Anticipating the behaviour of the other player can have a major impact on strategy choice, especially in iterated or repeated encounters between two players where previous choices may give an indication of future choices [89].

Individual player's utilities for different game outcomes are usually expressed as either a Tree Diagram or a Payoff Matrix. A Tree Diagram contains information about the order of play while in a Payoff Matrix it is assumed that each player moves simultaneously. A player is considered to be rational meaning that he chooses the strategy that maximises his utility, and a strategy is considered to be dominant if it yields the best outcome, regardless of the other player's move [89].

Strategies are in Nash Equilibria if they are the best response to each other,and an outcome is considered to be Pareto Optimal if there is no other outcome that could improve the utility for one player without reducing it for another [89]. Some games offer outcomes which can benefit all parties, others are strictly competitive or zero-sum, with a gain for one party being matched by an equal loss from another [206]. There are a variety of games of varying complexity that are used in modelling agent interactions. Three key games that are commonly used are:

- **Prisoners Dilemma** two players are said to be charged with the same crime. Each has the option to cooperate by maintaining silence or to defect by confessing their role and implicating the other. The payoff matrix for the game means that the maximum benefit any party can achieve is if they defect and the other cooperates, and vice versa. If both parties defect each receives a lower payoff than if both had cooperated [89].
- **Stag Hunt** Each party can choose to hunt a hare or a stag. If he chooses to hunt a stag, a player can only succeed if the other player makes the same choice, whereas each player is individually able to hunt a hare. The stag is worth more to each individual than the hare. The game differs from the Prisoners Dilemma in that the maximum payoff results from mutual cooperation, rather than individually defecting while the other party cooperates [89].
- **Chicken** Two players are driving towards each other. The highest payoff for either player is if they stay on the road while the other party backs down and steers away. The lowest payoff for each party

is if neither steers away. If both players steer away then there is a medium payoff [206].

2.3.7.3 Influentials

Influentials are a feature of the social networks that are used in the simulations presented in later chapters of this thesis. netThe *two step flow* theory of information transfer suggests that influence flows from the mass media, through opinion leaders, to their followers [201]. Under the two-step flow theory, influentials are not the heads of formal organizations, for example the editors of newspapers, but are private individuals who have a high-degree of personal influence. The theory does not explicitly specify the mechanism by which influence is transmitted, but holds that the role of influentials is important to the formation of public opinion [35]. The role of influence has been studied in the context of online social networks where electronic traces make information transmission transparent [190, 35].

Watts and Dodds [201]create a model of the opinion formation process in which each agent has to make a binary decision about an issue (0 or 1). The agents are arranged in a network, with each having an influence drawn from a random distribution. Agents have undirected links, so can be influenced, as well as exerting influence. The agent's influence is a function of the number of connections it has, and its authority on the topic. Agents are heterogeneous in the level of influence they exert, but the distribution of influence is relatively centred around the average any agent is unlikely to be many times more influential than any other agent. The relative number of other linked agents favouring a particular choice will influence the decision of each individual agent. Each agent has a threshold when sufficient other linked agents have chosen a particular outcome and the threshold is reached, the choice switches from 0 to 1. Initially all individuals are in decision state 0, except one who is activated. This activation may create subsequent activations forming a cascade.

Watts and Dodds distinguish between local and global cascades, with global cascades occurring only when a critical mass is reached amongst the early adopters — as agents who are activated after being influenced by just one neighbour. They compare the average size of cascade started by an influential, to the average size create by an average member of the population, under different network density conditions. They find that the development of cascades is more dependent on the overall network structure than on the influence of any agent, but also that influentials are more likely to trigger larger cascades.

Apart from varying the density of the network structure, they also test variations of the network topology by introducing groups. They find that the role of influentials is even less important in non-random, group based networks, possibly because influentials are more likely to group with other influentials who are more difficult to influence. They also experiment with the influence mechanism introducing a model in which an agent can be in a number of finite states *susceptible* (S), *infected* (I), *recovered* (R) model. Agents can be infected with probability beta on an encounter with an infected agent, and recover at a rate r. Under this mechanism, every contact between an I agent and an S agent is independent of any other. Watts and Dodds find that under the SIR model, the role of influentials remains limited [201].

2.3.8 Complexity

There is no universally agreed definition of complexity. Shalizi defines a complex system as ‘one with many parts, whose behaviours are both highly variable and strongly dependent on the behaviour of the other parts’ — thus excluding systems whose parts are independent of each other [181]. Similarly there is no agreed definition of emergence, though most definitions of an emergent entity agree that it is a higher order structure that emerges from a lower order system, usually due to local interactions between the individual components of the system. For Gilbert [75] ‘ A phenomenon is emergent if it requires new categories to describe it which are not required to describe the behaviour of the underlying components’. Epstein defines emergence as ‘stable macroscopic patterns arising from local interaction of agents’ [58].

Alongside the different definitions of emergence there are also varying views of how emergence can be broken down into sub-categories, including nominal, weak and strong emergence. Phan and Amblard consider that the philosophical literature largely agrees on a definition of strong emergence as demonstrating downward causation and irreducibility. But they conclude that there is little agreement on how weak emergence should be defined [162]. Despite the lack of agreement, the different perspectives and definitions proposed by different authors are informative in offering illuminating perspectives on the underlying processes.

Epstein puts the modern definitions of emergence in a historical context. The study of emergence began with the British Emergentists in the 1920s with Broad and his book ‘The Mind and its Place in Nature’ playing a central part. But Epstein argues that because the British Emergentists believed in the irreducibility of the emergent whole, contemporary complexity research should distance itself from the movement. He describes the British Emergentists approach as anti-scientific, because of their belief that the emergent system could not be explained in terms of the interactions of its parts. In his view emergence should only refer to something that does not preclude explanation [58].

Muller puts an emphasis on the role of observation, and the particular perspective from which the emergent process is observed. He argues that for a state to be emergent it must be interpreted as such by either an external observer, in which case the emergence is considered weak, or by the entities themselves, in which case it is considered strong. As an example of what he considers weak emergence he cites the movement of ants transporting food. The ants are following pheromones but the observer identifies a global phenomenon – a path. As an example of strong emergence he cites a collectively produced institution which constrains the actions of agents [151]. Phan and Amblard point out that an agent’s cognitive ability impacts his capacity to recognise these emergent institutions and so in turn the viability of downward causation and its resulting patterns [162].

Conte [42] examines emergence in the context of directions — presenting three ways in which a process can act. She describes vertical emergence as a process that occurs at a level of analysis which is either higher or lower than that at which the generating conditions are operating. Under this categorisation, bottomup emergence refers to macro level formations resulting from micro-level actions, such as the creation of coalitions or collectives. She refers to top-down emergence as the impact of social context on the individual, in terms of both cognition and performance. She cites reputation as an example of top-down emergence and identifies horizontal emergence as a change in an agent's behaviour to correspond to the expectations of another agent. She suggests that circular emergence is a process which starts at one level of analysis and proceeds to another, before feeding back to the first [42].

Gilbert [73] uses Schelling's model of segregation to illustrate varieties of emergence. In the model, agents rate their happiness based on the proportion of similar neighbours, with agents moving to a new location if the similarity proportion is below a threshold. Any threshold above 30 percent leads to the emergence of clusters of similar agents. Gilbert uses an amendment to the simple model to illustrate what he calls downward causation he introduces a crime rate, which is determined by individual action but in turn affects the ability of agents to move around the grid. He also introduces what he calls second order emergence the agent's recognition and response to emergent organisations [73].

Bedau [20] identifies a distinction between emergent and resultant properties, in which resultant properties can be explained from the properties of the components. He then describes 3 types of emergence, nominal, weak and strong, and points out that a macro level in one context might be a micro level in another for example a cell emerging from biomolecules and an organism emerging from cells. He defines nominal emergence as being a macro property that can not be a micro property, for example the molecules in a cup of water do not individually have the property of fluidity. He describes strong emergence as starting where scientific explanation ends — having irreducible causal powers at a macro and micro level [20].

Bedau identifies weak emergence as sitting between these two extremes something that is derivable from the underlying micro component's facts, but only by simulation. He identifies The Game of Life as an example of a process displaying weakly emergent behaviour [20]. Baker argues that although the objectiveness of such a definition is useful, it relies in turn on a clear definition of simulation, and its capabilities. In particular, it requires that no analytic solutions or shortcuts can be found to achieve the same result [13]

2.3.9 Trust and Cooperation

Trust is widely considered to facilitate cooperation [22], which in turn allows agents to interact and collectively achieve their goals. An agent's trusting disposition towards another agent increases the likelihood active cooperation, for example an online healthcare referral service may present a choice of trusted suppliers with which an agent may proceed to cooperatively interact. For Birk, trust is an objective measure of the desirability of interacting with that agent [24]. Ramchurn *et al.* define trust as a belief that an agent will do what it says, reciprocating for the common good of both rather than defecting to enjoy a higher payoff. They argue that a trusting relationship is equivalent to a cooperative strategy [165].

Ramchurn *et al.* identify two broad types of trust in a multiagent system — individual level and system level. At an individual level, an agent needs to be able to gather information and reason about the trustworthiness of another agent. This information could come from its own experience, or the reported experiences of others. They classify individual trust as either learnt (evolved), based on reputation, or socio-cognitive. In addition they identify a system-level trust which relates to the trustworthiness of the transaction protocols used in an interaction. They point out that some mechanisms, like the English auction, actively encourage trustworthiness on the part of the participants [165]. Whereas in a sealed bid auction there is a possibility for the parties to lie about their bids, in an English auction all bids are transparent.

2.3.9.1 System Level Trust

Ramchurn *et al.* argue that there are specific interaction protocols that can alter the extent to which a transaction can be subverted by the participant's incentive to defect. These can either be brought about by reducing the participant's utility for lying or colluding, by spreading each participants reputation, or by requiring endorsement by another party. They suggest that English and Dutch style auctions are more likely to cultivate honesty on the part of the auctioneer because the bids are publicly displayed, whereas in sealed bids they are not. They believe that most interaction protocols are nonetheless vulnerable to strategic lying and collusion between agents [165].

2.3.9.2 Different Types of Reciprocity

Berg *et al.* argue that a fundamental assumption in economics and many other social sciences is that individuals act in their own self-interest [22]. In open distributed computer systems different entities represent different stakeholders and have their own aims and objectives which they wish to achieve, and therefore the most obvious strategy is to maximise their own utility [165]. Similarly in biology, the theory of evolution rests on the notion of competition between individuals who promote their own success at the expense of their competitors [155]. Despite this, cooperation is widely observed in real-life situations, with humans exhibiting particularly complex forms of direct and indirect cooperation [156]. Axelrod and Hamilton suggest that in general, benefits are disproportionately available to groups of cooperating individuals, but that the persistent problem is that although individuals can benefit by mutual cooperation they can benefit more by exploiting the cooperation of others [10]. Nowak [155]developed a five part classification of the possible causes of reciprocity and cooperation in different situations.

- **Kin Selection** – individuals are likely to favour cooperation with genetic relatives.
- **Direct Reciprocity** – in the real world there is also cooperation between unrelated individuals, Direct reciprocity seeks to explain cooperation under the assumption of repeated encounters between the same individuals. Nowak uses the iterated Prisoners Dilemma to illustrate that in repeated encounters, cooperation can appear through the use of certain strategies such as *win-stay, lose-shift* – repeating your previous move if you're doing well but changing it if you're not.
- **Indirect Reciprocity** – built on reputation, this is relevant to situations in which one individual is able to help another but direct reciprocation is impossible. Nowak argues that indirect reciprocity has

significant cognitive demands requiring the ability to remember our own interactions, monitor the wider group and communicate reputations to other individuals.

- **Network Reciprocity** – this takes account of the fact that individuals are not equally likely to interact with all other individuals due to spatial or social network considerations. These limitations mean that some individuals interact with each other more than others, and mean that co-operators can form semi-isolated local clusters in which they help each other.

- **Group Selection** – conceptualises natural selection as working at multiple levels at the individual and the group level. Groups which contain more co-operators are more likely to be successful. Although defectors within the individual groups are more successful, groups that contain more co-operators do better than groups that do not are more likely to procreate to the extent that the group splits into multiple groups.

2.3.10 Evolution of Cooperation

A commonly cited example of how cooperation can emerge in a multi-agent system is based on an iterated version of the Prisoners Dilemma game. Axelrod and Hamilton use the game to explore the evolution and maintenance of reciprocal cooperation. They distinguish a game played between individuals who are unlikely to meet again from a game played by individuals who are. They argue that because the dominant strategy is to defect, a society based on one-shot, non-repeated interactions will evolve into a stable state made up of a collection of defecting individuals. They argue that a viable candidate strategy for explaining the evolution of cooperation should be able to precipitate movement from a non-cooperative to a cooperative state, and also support a stable cooperating regime. They define an evolutionarily stable strategy as one which, if used by a population of individuals, can not be invaded by a mutant individual adopting a different strategy.

Following a competition between a variety of strategies submitted by contestants, they found that the strategy of Tit-for-Tat was most successful. The strategy starts by cooperating on its first move, then does whatever the other player did last time, remembering only one previous move [10]. Ramchurn argues that not all multi-agent interactions are competitive. In some scenarios an agent may be self-interested but achieve a payoff from the success of the overall group [165].

2.3.11 Reputation

Reputation, defined broadly by Ramchurn as the opinion or view about someone on something [165], can help an agent to evaluate another's trustworthiness in a context of *one-shot* interactions — i.e. where there is not necessarily a history of repeated interactions between any two agents [101]. A well-known example of reputation management is the eBay feedback facility, through which buyers rate sellers with a +1 or -1 score, providing a guide for future buyers about their reliability. Nowak and Sigmund point out that one-shot interactions are becoming increasingly frequent in anonymous global markets, replacing the type of long-lasting, repeated exchanges that may have happened between members of the same village [156].

A functioning reputation management system can aid the development of a system of indirect reciprocity in which individuals are willing to help one another even though direct reciprocation is impossible. Ramchurn *et al.* split the requirements of a reputation management into three parts gathering ratings from the agent community, aggregating the ratings into a framework that facilitates reasoning and promoting ratings that accurately reflect trustworthiness [165].

Nowak argues that indirect reciprocity has significant cognitive demands requiring the ability to remember our own interactions, monitor the wider group and communicate reputations to other individuals [155]. Janssen identifies a number of problems with reputation systems in e-commerce, including the difficulty of aggregating positive and negative feedback into a meaningful score, barriers to entry for those with no reputation, and a lack of incentive for self-interested agents to provide feedback [101]. Ramchurn *et al.* suggest a number of ways that individuals can be incentivised to provide feedback, such as receiving discounts in future transactions [165]. They also point out a further problem, the possibility that agents may exit a reputation system, leaving behind a reputation, and then re-enter it under a new identity.

2.3.12 Cooperation between Similar Parties

2.3.12.1 Tagging

Tags can be used to represent a socially distinguishable mark or signal that indicates that an agent exhibits a particular behaviour, or purports to exhibit that kind of behaviour. In many applications the semiotic connection between the sign and the signified is allowed to be fallible — an agent that does not actually exhibit these characteristics can falsely adopt the tag. Kim identifies two broad ways in which tagging is used in models. In the first, the *action strategy*, an agent identifies what action to take based on a tag. This does not affect the probability of interaction agents may interact randomly. In the second, the *selection strategy*, the tag affects the probability of interaction, with agents more or less likely to interact based on a tag value [113].

In applications that involve cooperation between agents, tags allow classes of individual to recognise each other, developing cooperation through clustering and interacting with similarly tagged individuals [53]. Riolo *et al.* [168] use tagging in the context of a selection strategy to show how cooperative sub-groups can achieve short term domination through selective reciprocation between similar entities. In their model of 100 agents, each agent is given a tag, with a value between 0 and 1. In addition, each agent is given a level of tolerance so that the tag value of any other agent is either within their range of tolerance or outside of it. In each turn, each agent evaluates the tag of one other random agent. If the evaluated agent's tag is within the tolerance of the evaluating agent, the evaluating agent makes a donation to the evaluated agent, with the donation benefiting the recipient more than it costs the donor. If a donation is made, it adds or subtracts from each agent's score variable.

After three turns each agent compares its score with the score of another randomly chosen agent. Of the two compared agents, the one with the highest score procreates. By default the resulting, procreated agent inherits its parent's tag and tolerance. However it also has a 10 percent chance of mutating its tag to another random value between 0 and 1 and a 10 percent chance of adding a mean 0, standard deviation 0.01 random normal number to its tolerance threshold. In the event that the tolerance threshold falls below 0 it is reset to 0.

The model shows that a group of agents with similar tags tend to form a dominant cluster that donate and receive within their group at a high rate. The dominant grouping is vulnerable to exploitation by mutant offspring who fall within the tolerance range of dominant group members, but themselves have a low tolerance. These new group members are therefore likely to receive donations but not to reciprocate. Because they garner high scores they are likely to procreate, and as a result form a new grouping with identical tags and a narrow tolerance range. Members of this group are unlikely to donate to the existing dominant group, whilst receiving donations from it, and so the dominant group is replaced. Riolo *et al.* point out that the procreation mechanism could also be interpreted as a learning mechanism, in which the same agents adopt the tags and tolerances of their more successful pairs. Kim highlights some of the key dependencies in the Riolo *et al.* model including the fact that high levels of cooperation are possible because it allows a *less than or equal to* rather than a *strictly less than* tolerance criterion meaning that an agent with tolerance 0 is able to cooperate with identical agents [113].

Edmonds *et al.* give an example of an application in which a similar mechanism is used — a peer to peer file-sharing mechanism, similar to BitTorrent. Each machine is a client and server and interacts in a directed network with a limited number of other known machines, rather than the whole network. Each node is able to search beyond its local grouping by sending queries to its contacts, which in turn pass them to their contacts up to a maximum number of degrees. If a node has a file and is sharing it sends it to the originator and the originator node's satisfaction level increases. Each node's satisfaction level decays with each cycle down to a threshold, at which point it copies the strategy and connections of a better performing neighbour. Their simulated system leads to the formation of a core interconnected partition with peripheral branches feeding into it. Isolated groups form and die relatively quickly. They find that co-operators who share files have higher satisfaction levels than defectors who do not [53].

2.3.13 Norms

Wooldridge [206] defines a norm or social law as being an 'established, expected pattern of behaviour'. Some examples of norms in human societies include fashion, social etiquette, and driving practices. Andrighetto *et al.* [8] argue that norms are distinct from values and other cultural products in the sense that they are enforced by sanctions. They explain the motivation to enforce such sanctions as arising from the fact that norms are public goods whose benefits can be enjoyed by all of a society, so non-adopters are therefore liable to be considered free-riders by the other members. This enforcement and defence of norms is widely considered to be key to their transmission. Sen and Airau list the main features of conformity to norms as being reduced social friction, a lower cognitive burden, and easier coordination [178].

Epstein [58] points out that norms provide a reduced cognitive burden, by constraining the range of choices that an agent faces. He argues that a norm becomes a pattern of behaviour that, once adopted, is enacted without thought. He also suggests that the individual computing required in considering its adoption is inversely proportional to the strength of the norm. In addition, he proposes that agents not only learn how to behave, but learn how much they need to think about how to behave. Epstein illustrates the cognitive burden of uncertainty using an agent based model in which agents are required to make a more detailed evaluation of norms when there is more diversity around them. Andrighetto *et al.* [8] argue though that norms are not just constraining, they can also add new goals to an agent's repertoire that it may not have otherwise considered.

Castelfranchi *et al.* [33] argue that the action constraint approach is pervasive in the Artificial Intelligence literature, which primarily considers the role of norms to be 'norms of coordination. In constraining the repertoire of actions available to an agent, this in turn reduces the actions that the overall system can perform. They argue that norms of coordination are only a subset, that other tendencies should also be included. To illustrate this they build a simulation in which norms are used to reduce aggression. Agents move around a grid, using a set of routines that include eating, pausing, attacking another agent, moving randomly and moving towards food. The tendency to show aggression is exhibited through the likelihood of using the attack routine, and different methods of consideration can be given to whether or not it should be applied, including blind (unthinking aggression), strategic (attack if advantageous) and normative (refrain from attacking agents eating food from within their own territory). Overall, the study finds that with competing groups of agents using different strategies, the normative approach is less successful than the strategic approach.

Wooldridge [206]identifies two ways in which social conventions can be introduced into agent society. They can be:

- **Designed and imposed from the top down** – can be imposed with greater control but not all of the characteristics are known at outset.
- **Emergent from within system** – allows dynamically developing norms to emerge to address things unknown at design time.

The emergence of norms requires a mechanism through which a convention can become established agents need to be able to monitor the behaviour of other agents, recognise norms as being such, and be able to modify their own behaviour. Wooldridge gives an example of a scenario in which a group of agents have a choice of t shirts to wear, and a goal of all wearing the same shirt. He discusses a number of update functions which an agent could use, for example a simple majority, or highest payoff [206].

Sen and Airiau [178] develop a simulation based on the emergence of road traffic norms. Their approach differs from many norm based simulations in that interactions are between pairs of agents, agents have no visibility of the wider group but learn normative behaviour through random, anonymous, one-shot pairings. This means that agents cannot adapt to the specific characteristics of the other player as they might in an iterated game between two parties. The pay-off that each agent receives is based on the outcome of each encounter.

Sen and Airiau develop a model in which two drivers face a social dilemma — simultaneously arriving at an intersection from different roads. The payoff matrix is such that for each player, going at the same time as the other is the worst outcome, followed by both yielding, followed by yielding while the other goes and finally going while the other yields. Sen and Airiau see an efficient outcome as being a norm such as 'yield to the driver on the right'. Each driver is randomly assigned to its arrival position so will either have a car on its left or right. Each driver has a learning algorithm which it uses to develop a policy to deal with each possibility. The learning algorithm is assigned from a choice of three and is consistent to the agent throughout the game. They find that in any simulation with three or more agents, a norm emerges. Following 1000 runs, the split between yielding to the left and yielding to the right was approximately equal. They argue that this supports the notion that private experience can lead to norm emergence, without knowledge of non-local interactions.

Luck *et al.* [134] distinguish between the processes of adoption and compliance with norms. They suggest that there may be cases in which automatic compliance is counter-productive, since an agent may have a personal goal which conflicts with the norm. In these cases, an agent may choose not to adopt a norm, making a choice between its own goals and the requirements of the norm, with the implicit play-off between the stability of the society and the agent's own objectives. In these cases, Luck *et al.* suggest that in order to achieve compliance — for social norms to dominate individual aims, there is a need to apply punishments and rewards. In Sen and Airiau's example, the motivation to comply with the norm is expressed through the pay-off matrix a higher pay-off could be considered a reward, and a lower pay-off a punishment [178].

2.3.14 Randomness

Randomness is widely used in agent based modelling. Models that contain randomness differ from deterministic models, which display completely predictable behaviour, in that there is some uncertainty about the outcome [72]. Epstein [58] argues that in a deterministic model without any random element, every state in the process follows deterministically from the state before, and so the final outcome of any model can be calculated from initial conditions. This sensitivity to initial conditions and the model's resulting deterministic development and outcome is also referred to as path-dependence [76].

There are a number of ways in which randomness can be introduced to alter a model's development. One example is by introducing errors into communication between agents [74]. In other models, like Riolo *et al*'s model of tag-based cooperation [168] the random mutation of certain agent characteristics is fundamental to the developments of the models dynamics. Another reason modellers introduce randomness is to substitute for real information for example links may be randomly allocated between agents when there is insufficient empirical information to establish the real relationships in a network [74]. Epstein [58] points out that even when stochasticity is introduced into a model, since randomness for a computer is produced by a deterministic random number generator, a model can be recreated by using the same seed for the generator.

2.4 Model Selection and Agent Based Models

Model selection and validation methods are central to the model and theory selection process which is proposed. In this section existing work on model validation is reviewed, particularly in the context of agent based models. Some of the issues involved in establishing causality through modelling are then considered.

2.4.1 Validation, Verification and Selection of Agent Based Models

It is widely agreed that although verification and validation are important issues in agent based modelling, there has been a lack of accepted methods and protocols to consistently achieve them. This has led in turn to a perception that there is a lack of robustness in the method, and generated a reluctance to adopt it amongst mainstream economists and other potential users [205, 140, 18]. The issue is compounded by the fact that some characteristics of agent based models, including feedback loops, emergent properties and adaptive behaviours, can make the validation of models problematic [150]. In some cases, there is uncertainty on the part of the modeller about what outputs these kinds of feature might create, meaning that it may be hard for them to know if an unexpected outcome is a novel feature or an error.

Although the terms verification and validation are sometimes used interchangeably, most authors distinguish between the two activities. In general, the term verification is used to refer to issues around accurately implementing the specifications of the model in a computer program, including conceptualisation and programming [70]. Validation is usually used to describe the practice of making sure that the implemented model is a reliable representation of the target process [18].

Windrum *et al.* [205] identify four ways in which validation issues have contributed to the slow take-up of the method compared to others. They contrast the practices of agent based modelling to those of what they call *neoclassical* modelling essentially mathematical statistics. Firstly, they suggest that whilst the neoclassical modelling community has concentrated on a number of core models and sought to apply them to different research areas, the range of agent based models used is extremely diverse and unfocused. Secondly, there has been little structured comparison of how different models perform — they have been used in a wide range of applications but there has not been sufficient evaluation and comparison of how different models perform in the same area. Thirdly, they find that there is a lack of protocols for building and evaluating models. Finally, like other authors they report a general uncertainty amongst practitioners about how agent based models should relate to empirical data and whether they should be validated at all [205, 3, 137, 192].

2.4.2 Verification

Galan *et al.* [70] identify two general types of issue that should be monitored as part of a verification process — one being errors and the other artefacts. They describe an error as a mismatch between what the developer thinks is happening and what is really happening, for example looping through a subset of agents when he intends to loop through all of them. They consider an artefact to be an aspect of the program that is considered to be unimportant but takes on more importance when the model is run. As an example they suggest the possibilty that the number of grid cells is chosen arbitrarily for a particular model, but that when the number is changed it is found that the key result is no longer present.

Different types of error and artefact can be introduced into the model at different stages in the development process. Galan *et al.* [70] cite Drogoul *et al*'s (2003) definition of three conceptual roles in constructing an agent based model: the thematician, the modeller and the computer scientist. They identify problems that may occur within each role and at the interchanges between them. The thematician defines the objectives of the model and conceptualises it he may create errors but can not create artefacts since he is dealing with the intended features of the model. The modeller transforms the requirements of the thematician into a specification for the computer scientist. The modeller can create artefacts because he will have to add in additional assumptions to make the model viable. Galan *et al.* consider it unlikely that the modeller would create errors. The computer scientist finds an approximation of the model and programs it he creates errors and artefacts. There are many methods of verification proposed in the literature, including:

- **Dynamic testing** — Amblard *et al.* [3] argue that model verification goes beyond checking for bugs and requires running the model and checking for unexpected behaviour.

- **Code Inspection** — Galan *et al.* [70] point out that its not always the case that anybody knows what the computer code is doing even the programmer. The process involves experts paraphrase the meaning of each line by verbalising a summary [143].

- **Docking —** Another team attempts to recreate the model, then compares the output with the original in terms of their distributions, relationships and numerical similarity [70].

- **Software Verification** — This covers settings and methods which may not be transparent to the programmer, such as whether or not the default implementation is to use floating point arithmetic [70].

- **Assumption Checking** — Identifies the assumptions in the model — all simulations contain assumptions, some are known, others not [70].

- **Stability Analysis —** This identifies a model's tendency or otherwise to converge to equilibrium.

- **Sensitivity to Initial Conditions** — This is analysis of how initial conditions impact on the results of the model.

- **Sensitivity Analysis of the Parameters** — Implemented by systematically varying the parameter values and observing the impact on the model. Sensitivity analysis in agent-based model may be complicated by the non-linearity of the underlying functions, and the possible cross-dependency of the parameters. Ginot and Monod [78] define two types of sensitivity analysis local and global. Local analysis refers to the parameter values that are used in typical operation of the model, while global refers to a wider range that is not commonly used. Because the wider range of possible values is so large, they present several methods for reducing the number of models that need to be tested, including random-sampling of parameters, and a factorial experimental design to accommodate cross-interactions between variables.

- **Destructive Testing** — Midgley *et al.* [143] argue that destructive testing can be informative about the reasonable boundaries of the model. They suggest that by seeking to maximise certain objective functions, for example the level of error between the empirical and modelled data, they can test the sensitivities of a model and its likelihood of breaking down when faced with extreme parameter values or unusual combinations of parameters.

2.4.3 Validation

The type of validation and its success criteria depend to a large extent on the objectives of the model being reviewed. Marks [140] identifies some broad possible aims for simulation: to explain a phenomenon, to explore a phenomenon, or to predict the outcome of a phenomenon. Amblard *et al.* [3] refer to a further category, discussing the possibility that for some modellers it may be the case that the sole value of the model is in what is learnt in the process of building it. At one extreme very little validation is needed. For models built with the purpose of illustrating or understanding a process, correspondence with empirical data is only a secondary criterion [3]. At the other extreme, for example when a model is intended to be predictive, validation may be an important criterion for gaining credibility for its predictions.

Amblard *et al.* [3] argue that a necessary pre-requisite for validation is that the model can actually be validated. This requires that the model's results can be reproduced to a reasonable degree, which may involve using a multi-platform language to implement the model, and controlling the seeding of random-number generators.

Windrum *et al.* [205] consider three broad approaches to validation:

- **Indirect Estimation** — The term indirect estimation is used slightly differently by different authors. Chen *et al.* [36] use it to describe a general approach to calibrating a mathematically intractable model. Windrum *et al.* [205] use it to describe a more specific four step approach to calibration. In step one the modeller identifies an empirical dataset that they wish to explain. In the second step the modeller builds a model such that the microlevel behaviours are theoretically accurate. The approach does not directly use empirical data to set parameters. In the third step, the parameters in the model are calibrated against the empirical data. In the fourth step the modeller reviews the clash between the theoretical underlying behaviour and the estimated behaviour.

- **Werker-Brenner Approach** — Similar to Indirect Estimation but uses empirical parameters to calibrate the model and restrict the possible values that parameters can take.

- **History Friendly Approach** — Uses the specific, detailed historical case studies of an industry to initialise interactions, parameters and decision rules. They focus on what they call appreciative theories, which emphasise the practical expectation of the modeller, rather than the academic theory of the discipline. The approach uses backward induction to arrive at the correct parameters, structural assumptions and initial conditions that replicate the actual industry history. The method is data intensive [138].

2.4.4 General Issues in Empirical Validation

Windrum *et al.* [205] identify some areas which they argue are common to all types of empirical validation. They suggest that in general there are two sets of data involved in the validation process. On the one hand there is the empirically observed data which is the outcome of an unknown real world data generating process. On the other there is the modeller's representation of that process which attempts to produce a good approximation to the actual data. They argue that the neoclassical model does not necessarily aim to provide a causal account of the movements in the data — the intention is to find a function that recreates the shape rather than one that causally explains it. Explanations are sometimes retrospectively interpreted using the models inputs. This type of modelling is not particularly useful to explore alternative future scenarios because the causal factors that lead to different possible outcomes are unclear [3].

Amblard *et al.* [3] argue that the purpose of an agent based model is different to the neo-classical model in that it is concerned with trying to explain and understand the mechanisms that create the empirical data from the perspective of the individual. The modeller attempts to identify possible individual level behaviours that may have generated the empirical data.

Marks [140] presents a classification of conformity between the modelled and the actual variable that rates the model as useless in the event of no correspondence, through degrees of completeness and accuracy. Under his classification the best possible model is described as *complete*, in the sense that it captures all historical movements in the empirical data, and *accurate* in the sense that it does not demonstrate movements that are not present.

2.4.4.1 Calibration

The process of calibration aims to minimise the difference between the models outcome and the empirically observed data by adjusting the changeable characteristics of the model. In the context of an agent based model these include the following:

- **Initial conditions** The states and characteristics that the model is initialised with can make a significant difference to its development. Initial conditions may be manually defined, or allowed to develop using data from further back in time.

- **Agent behaviour** — Agents may respond to a certain situation by adopting a different rule rather than by varying a usage parameter. Moss and Edmonds give the example of changes to the rules governing financial markets after the 1929 crash [150]. Moss and Edmonds argue that agents are not necessarily implemented in a way that is a compatible with descriptions of the behaviour of observed entities but may be players in a game theory setting, as genetic algorithms [150].

- **Environmental factors** — Environmental elements of the model, such as the network structure of the links between agents, may be varied to improve the level of correspondence with empirical data [205].

- **Parameter Variation**— Chen *et al.* [36] identify three broad methods for calibrating the parameters in an agent based model. The first approach is a direct estimation which involves deriving a mathematically tractable version of the model and applying statistical methods directly to it. The second and third methods are *indirect*. The second involves creating aggregate series from a set of model runs and deriving the optimal parameters by comparing the runs with the actual data. He refers to this as the 'Method of Simulated Moments' and identifies one of the key issues as being the practical requirement of selecting the subset of parameters to test. The third approach is to evolve the model towards an objective function. Terano [194] proposes a version of this method that he calls *inverse simulation*. Rather than creating all possible models and comparing the outcomes against an objective, the process sets a macro-level objective function and attempts to evolve the simulated model to fit the objective. Midgley *et al*'s [143] supermarket model is an example of this approach, embedding the model in a genetic algorithm in order to minimise a distance function.

  As an alternative to inductively estimating the model's parameters, Garcia *et al.* [71] describe an approach in which each agent's parameters are based on the results of a conjoint study of real individuals. The agents take their values from the declared order of preferences for wine characteristics. Similarly, Sengupta and Glavin [179] base some of their agent's parameters on empirical purchasing data, although they indirectly estimate the weights for combining the parameters into an overall utility score.

  Macro level-studies, such as the Artificial Anasazi model [58, 102], aim to manipulate parameters to maximise fit against an expected outcome at an aggregate level, and is usually conducted against higher level aggregate patterns, often time series [149]. Several existing studies have successfully used learning methods to learn models of individual agent behaviour. Some of these have looked at the actions of the individual in isolation of other agents, for example the Letizia system was developed to model an internet user's browsing behaviour in order to suggest future actions that the user might take. It made these predictions based on classifying the activities that the user performs, such as bookmarking a page, and following links, and through this inferring the importance that the user place on each site [128]. In a banking context, Chiang *et al.* use a goal based search method to search for customers who are likely to move to another bank, based on identifying a sequence of banking transactions that indicate their intention, but [37].

From a group modelling perspective — learning individual behaviours in the presence of interaction with other agents, Lee *et al.* learn the parameters of an agent based model of individual and group pedestrian behaviour using a video capture system with vision based tracking to monitor the motion of individuals in a crowd. They aim to identify each agent's perceptual state and their response to it, learning the model for each individual agent based on the low-level behaviours. The state that each agent is in at any point reflects the motion of nearby agents, its own motion, and environment features. Group behaviour is formed through the agent's sensitivity to its environment. In each state the individual has a range of possible actions [125].

A core part of the model's capability is the ability to predict how an agent's behaviour might change when environmental circumstances change, for example when an obstacle is inserted or removed [15]. Gillies *et al.* [77] use a similar approach in calibrating the behavioural parameters of a game AI character to reflect the responses and movements of a real actor to stimuli in the environment. The data they used was collected from video capture of the actor's movements and interpreted using principal components and clustering methods. The interpreted movements were combined with a corresponding dataset of events in the environment that the actor was considered to be responding to. The calibration process involved relating the transitions between a number of states that the character may be in with the external,environmental events. When the character receives a specific input he is more likely to transition to a certain state in the next time period. The environment may constrain the set of choices that can be made, provide information on which action selection decisions may be made, and influence state transition probabilities. The physical environment may also impose constraints concerning which areas are accessible to an agent, by what means the agent is able to move around, and the degree to which an agent can modify its surroundings [157]. Bandini *et al.* [17] give an example of a model in which agents are particles which are subject to and also generate forces. In this case, they argue, the environment determines the overall dynamics of the system. Once parameterised and transferred into an ABM, the environment is the virtual world in which the agents act [74]and provides the conditions under which an agent can exist [157].

Approaches to learning action selection mechanisms depend partly on how many choices there are — if there is only one choice, for example whether the agent should act or not, then binary choice models can be used. If there are multiple discrete choices then multinomial discrete choice models. In a transport modelling context, Antonini *et al.* [9] use a discrete choice model to learn the action selection behaviour of pedestrians, taking into account their interactions with other individuals. Their data is based on video of crowds moving around the entrance of a metro station. They assume that the destination is already known. Each pedestrian is modelled in terms of their choice of speed and direction, which are banded into discrete categories [9].

2.4.4.2 Validation at Different Levels of Aggregation

In many agent based models it is necessary to validate the accuracy of model outcomes at multiple levels of aggregation. Moss and Edmonds [150] argue that validation should be conducted at micro and macro levels, although there may be other levels of aggregation in between. Micro validation refers to the behaviour of the individual agents comparing each agent against empirical data or an expected outcome. Macro validation is usually against higher level aggregate patterns, often time series. Sengupta and Glavin [179] provide a good example of model validation at different levels. Based on observed panel data over one year, they create a model of consumer purchase behaviour in the fruit juices category. They use part of the data to initialise the agent preferences, part of the data to calibrate their utilities, and the remainder to validate their predictions, taking into account actual changes in price, distribution and other factors. In their analysis, the micro-validation phase compares the actual product choices made by each consumer against the model's predictions with the outcome that 35 percent of product choices were accurately predicted. The macro validation phase compares the overall market share of each product derived from the model with the empirically observed share, finding that the correlation coefficient varies between 0.43 and 0.57.

Page [158] examines several examples of *aggregation failure*, which occurs when patterns which are visible at micro level can not be detected at macro level. For example, what may appear to be a stable equilibrium at aggregate level can actually be a cycle of states at individual level. Page gives an example of a simple Voter model with six agents arranged in a circle, each having two neighbours. Each agent is initialised with one of two states, which he will change if both neighbours are in the other state. The overall result is a permanent *blinking*. At agent level there is permanent volatility, with each agents cycling through a different state in each time period. Observed from an aggregate level, the count of agents in each state remains constant over time. Marks [140] suggests that in some types of model, macro-behaviour is relatively insensitive to the specific actions of agents. This may means that the overall macro outcome is less complex than the micro-level behaviour that constitutes it.

2.4.4.3 Model Complexity

Midgley *et al.* [143] point out that, amongst other causes, in-depth knowledge of a domain area tends to lead a researcher to include nuances and sophistication that may increase the number of parameters to an impractical level. Models with too many parameters are subject to some specific problems:

- **Identification** — The problem of over-parameterisation refers to the inclusion of too many parameters in a model, given the amount of data available to validate it. A large number of parameters, relative to the validation data, leads to a model which can produce identical outcomes with different parameter values. In the worst case, over-parameterisation can lead to a situation in which a model can create almost any outcome [3]. In general, it is not possible for statistical inference to decide between different candidate models when the output of each is the same or very similar.

- **Complexity** — Marks [140] argues that emergent properties make a model more difficult to validate, and considers ways in which the complexity of a model can be quantified. He suggests that some models, such as Schelling's segregation model, would be difficult to validate, even if the right data was available.
- **Analytical Tractability** — Scientific modellers usually recognise that it is impossible to model the full complexity of the world, and aim to try and isolate the main variables thought to reflect a phenomenon. In most models there is a trade-off between analytical tractability and descriptive accuracy. Higher levels of detail and more parameters lead to models that can not be solved analytically. More abstract and simplified models are more likely to be tractable.
- **Computation** — Terano [194] points out that from a practical perspective, models with many parameters create a computational requirement that can quickly become very large. For example, a set of ten parameters, with each one able to take ten possible values, can create ten billion permutations. Since it is hard to anticipate the dynamics of an agent-based model without executing it, at least one model run is necessary for each permutation of the parameters. If there are stochastic elements in the model then multiple runs may be necessary to calculate a stable average [32]. Ginot [78] proposes that time issues can be overcome to some extent by paralellisation of the model across processors. The requirement for testing multiple parameters can be limited by imposing a priori assumptions about the range of possible values is that a parameter can take [78]. Windrum *et al.* [205] argue that the imposition of assumptions is unavoidable in any scientific discipline.

Models that are too simplistic or contain too few variables are also susceptible to specific problems:

- **Omitted variable bias** — Taber [192] points out that the pursuit of parsimony should not be an objective in its own right, since omitting the wrong variables may lead to models that are causally invalid. If the right variables are selected the variance that is due to omitted variables and so left unexplained will become random disturbance around the error term — the difference between the model and the actual data. But if a variable is omitted that is highly related to the actual variable it will create systematic rather than random deviation from the error term which will lead to biased parameter estimates for the remaining parameters.

Although they are rarely mentioned in the agent literature there, are several well other defined problems in mathematical economics which are also relevant to agent based modelling. These include [84]:

- **Statistical Significance** — The modeller requires a method for judging the likelihood that the correspondence between the model and the outcome, and therefore the contributing role of the parameters being estimated, is not due to chance. In mathematical statistics this is usually described in terms of statistical tests such as the t test, F test and chi-square test.

- **Multi-collinearity** — Highly correlated input variables make it difficult to identify the independent impacts of the individual inputs.
- **Spurious correlation** — The possibility of mistakenly interpreting the relationship between two variables as being causally related.
- **Non-constant parameters** — Changes over time in the structure or *regime* of the model may mean that the parameters that agents actually exhibit vary over time.
- **Sample sizes** — Limited sample sizes may result in sampling artefacts being interpreted as real events in the underlying data generating process [93, 84].

2.4.4.4 Statistical Properties of Agent Based Data Series

Huckfeldt *et al.* [97] define an equilibrium value as a steady state that remains constant over time, although a fundamental change may occur which moves the process to a new equilibrium. The presence of an equilibrium does not mean that there are no underlying dynamics, just that the net result is balanced [97]. Huckfeldt *et al.* [97] cite the example of political support, which may stay constant for a particular party, even though there have been a succession of defections and new recruits. Epstein [58] argues that there are at least three scenarios in which an equilibrium will not be achieved the phenomenon is a non-equilibrium dynamic, the time-scales over which equilibrium might be reached are unrealistic, equilibrium exists but is unattainable. The attainment of equilibrium assumes some mechanism for doing so. Some agent based models create data which tends towards instability and extreme values. Moss and Edmonds [150] argue that the characteristics of data generated by interacting agents is more likely to be prone to out-of-equilibrium volatile episodes, with large numbers of values that are far from the mean. In addition, they argue that neoclassical assumptions that the data generating process follows a normal or binomial distribution are not reflected in reality. They suggest that as well as volatility, the process is likely to lead to an undefined variance and mean. Therefore social institutions would not produce numerical data as if it was drawn from an underlying numerical distribution. Amillon [4] reports similar behaviour in data from an artificial stock market. Because the trading agents are constantly updating their strategies, there is no single equilibrium that emerges from the model.

2.4.4.5 Causality

It is widely accepted that correlation, or association, does not imply causality, but that causality does imply some kind of association. Goldthorpe [80] proposes three conceptions of causality — causation as robust dependence, causation as consequential manipulation, and causation as generative process. Conte [41] argues that there the contribution of generativity in establishing causality can be seen in two ways: a weak thesis holds that since multiple paths may lead to the same result, the ability to generate something is necessary but not sufficient to explain it, while a strong thesis holds that it is necessary and sufficient.

The notion of robust dependence is used in equation based modelling to refer to methods that attempt to prove the causal relationship between variables. The fact that probability of Y increasing, given X is greater than the probability of Y given not-X does not necessarily confirm causality since both could be caused by a third element, Z [80]. Different methods have been developed that aim to move from correlation to establishing causality including Granger Causality, which tests the efficiency of X in forecasting Y [84]. Goldthorpe argues that this kind of predictability is not in itself enough but that what is needed predictability in accordance with a theory. He suggests that robust dependence can only ever be considered to be provisional at any time new data may become available that disrupts the data process which verifies robustness and reveals the relationship to be spurious.

Consequential manipulation uses experimental methods and is common in practical sciences like medicine and agriculture. In this context, causes are things that can be introduced as *treatments* in experiments — X is manipulated with all other variables controlled for and the outcome Y is systematically observed. Goldthorpe suggests that there is wide agreement that this is a stronger type of causality but also concern that it redefines the problem. Robust dependence starts with the effects and looks for causes whereas consequential manipulation starts with causes and looks for effects.

2.5 Theory Development and Selection

Theory plays a key role in the overall research objective — in particular the representation of a theory in terms of inter-relating agents and the ability to select a theory based on a given set of criteria. This section looks at some of the literature about theory as it relates to agent based models, steps in theory development, and different approaches to theory choice.

2.5.1 Theory and Agent Based Models

Livet [133] argues that the model and its design should be seen in terms of an overall knowledge framework, made up of three areas the empirical domain, the model domain and the conceptual domain. The empirical domain consists of observations of reality it is shaped by methods of data collection and the underlying world view of the observer. The model domain contains formal representations of what the observer wants to explain. The conceptual domain contains the theoretical background that contextualises the model. Livet suggests that in some applications, models are used as a tool for investigation with no basis in theory. Figure 2.2 shows how the different domains inter-relate.

Manzo [139] argues that as a consequence of its generative approach, the simulation modelling process emphasises the role of theory more than many other approaches, forcing the researcher to formalise a theory's specifications as they implement the model. Epstein [57] defines the generative approach as a process in which researchers *situate* an initial population of autonomous heterogeneous agents in a relevant spatial environment; allow them to interact according to simple local rules, and thereby generate — or *grow*

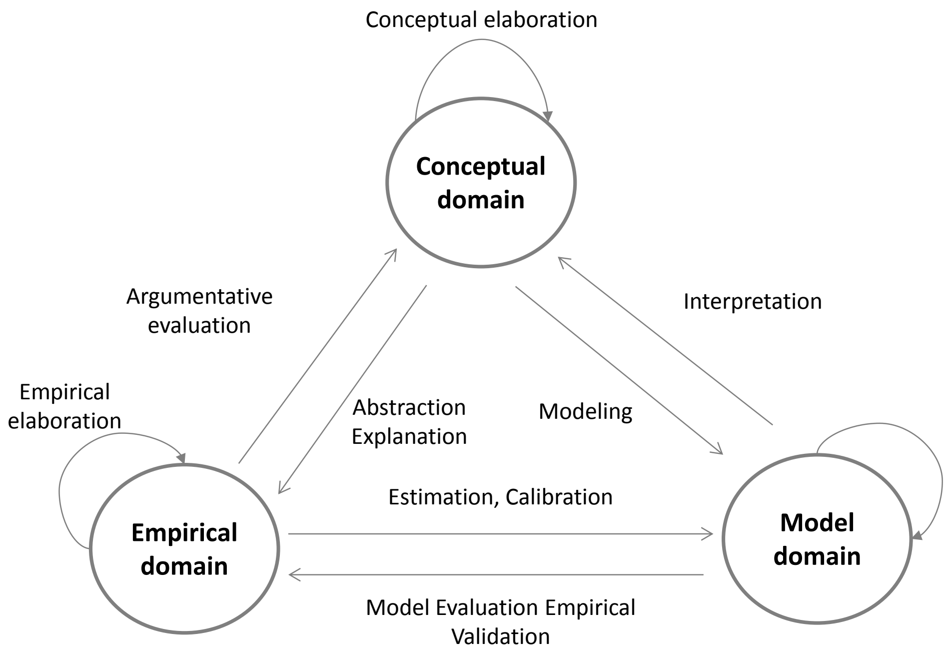

Figure 2.2: Overall knowledge framework — from [133]

– the macroscopic regularity from the bottom up.

Although Epstein and others make the claim of generative explanation for agent based modelling, similar claims are made [5, 123]for microsimulation models which do not feature interactions. Amstutz [5] argues that by building up from behavioural rules at individual level, microsimulation also has the potential to provide what he calls the *right answers*, and also explain the reasons for the answers. Amstutz [5] argues that by basing the relationship between inputs and outputs on a theory of behaviour, microsimulation allows a model to be used outside of the specific conditions that existed when it was estimated.

Conte [41] argues that even in the presence of a generative mechanism, there is still an important role for theory in interpreting the micro-level rules and macro-level outcomes in a model. She points out that in a model as simple as Schelling's model of segregation [176], there could be rival explanations of the individual level decision rules which change the interpretation of the model overall. For example, instead of being racist, the individuals may prefer homogenous environments.

Grune-Yanoff [85] argues that because many agent based models do not accurately reproduce the target state, there is no actual explanandum, so the things that they explain did not really happen. ABMs that do accurately recreate the target aim to represent possible causal histories. He contrasts Epstein's Anasazi model with a simulated model of a car-crash, suggesting that the car-crash model is based on laws that are proven to be true, while the behavioural rules in the simulation cannot be verified because different rules create similar fits. He suggests that supporting evidence should come from direct observation, well-confirmed theory, or results from externally valid behavioural experiments. He argues that there is little surviving theory to support the rules used in Epstein's simulation as being correct. He also suggests that it is unlikely that the rules used by agents are consistent through multiple social states. He suggests that because a model can only be a candidate explanation, and there is a vast pool of candidates a *filter* is needed to select possible causal histories through criteria that are independent from our evidence for certain causes, although he is unable to identify a method for selecting these criteria. He suggest that rather than providing a causal explanation agent based models provide at best a causal explanation. Elsenbroich points out that Grune-Yanoff's criticism of ABM is more a criticism of the data that sits behind the simulation than of the method itself [55].

2.5.1.1 Simulation and Theory Development

Davis *et al.* [44] argue that simulation is useful in theory development because it encourages precise specification of the problem, requires the boundaries of the theory to be clearly specified, and permits experimentation. They propose a 7 point roadmap for using simulation to develop theory:

- **Begin with a research question** — they suggest that this helps to focus the researcher and reduces complexity.
- **Identify a simple theory** they suggest a theory that is relatively undeveloped with a few constructs and related propositions

- **Choose a simulation approach** – the choice of approach depends on the fit of the research question. They review the benefits of a number of simulation methods and point out that the choice of the technique is important since it shapes the resultant outcome.

- **Create computational representation** — this involves implementing the chosen approach

- **Verify computational representation** — this includes sensitivity analysis, checking for internal validity and sense-checking the results against the theory, with the aim of ensuring that the model represents the underlying theory.

- **Experiment to build novel theory** — they suggest that this could include varying the varying the elements of the simulation, decomposing some elements into their constituent parts, varying assumptions, and adding new features.

- **Validate with empirical data** — this involves checking the results of the simulation against empirical data.

2.5.2 Theory Choice

Kuhn [119] discusses the role of quantitative and qualitative accuracy in choosing between theories. He argues that the history of science proves that in reality, theories cannot always be discriminated in terms of accuracy. He argues that his critics believe that theories can be selected on the basis of a set of objective criteria, but in fact two scientists committed to the same set of criteria can come to different conclusions, since the objective criteria chosen can be completed under the influence of subjective considerations. One theory may match better in one area and another in a different area, choosing between them means prioritising an area of coverage, which is not necessarily a scientific decision. In addition, scientists may make different theory selections due to their application of different weights to different criteria, or assumptions regarding different definitions — such as the definition of simplicity. Characteristics that vary from scientist to scientist can also make a difference, even when the scientific method is not in dispute. These difference in characteristics might include a scientist's previous experience, and the extent to which his current success is reliant on acceptance of a previous theory, or differences in personality such as risk aversion. He suggests that the history of science discards discussion of the merits of alternative theories which were not selected.

Kuhn [119] identifies five criteria for a good theory: it should be accurate deductions should be in keeping with reality; it should be consistent with itself and other accepted theories; it should have scope beyond its own domain; it should be simple; and it should reveal new research findings. Wacker [200] cites a number of characteristics exhibited by a good theory:

- **Uniqueness** — the theory can be differentiated from another, if two theories are identical they can be considered to be the same.

- **Conservatism** – the current theory cannot be replaced unless the new theory is better.

- **Generalizability** – the theory can be applied more widely than other theories, across multiple areas.
- **Fecundity** – the theory helps to expand the investigation into new areas.
- **Parsimony/ efficiency/ simplicity** – if theories are equal in all other respects then the simpler one is the best.
- **Internal Consistency** – the theory logically explains the relationships between the variables, and the relationships are logically compatible.
- **Empirical riskiness** – an empirical test of the theory runs the risk of refuting it.
- **Abstraction** – the theory is independent of time and space.

Kuhn argued that scientists see through the lens of their theory, which provides an interpretative framework and so pre-determines their perception and classification of entities. This leads to a situation of incommensurability between rival theories [88]. Because what is observed is observed through the lens of a theory, there is no neutral observation language — results from experiments cannot be compared because the words have different meanings, even when the same word is used [66]. Brewer and Lambert cite studies from psychology that provide evidence of episodic perceptual priming – the influence of previous experience on current perception. For example in one study, students were shown two balls falling one iron and one plastic, then asked to observe an experiment. The students who had initially hypothesised that the iron ball would fall at a faster rate were more likely to note this as the outcome of the experiment [26].

Brewer and Lambert also draw on studies in cognitive psychology to discuss the role of theory-ladenness beyond its influence on visual perception.

- **Attention** — they argue that it affects attention in the sense that scientists are more likely to pay attention to observations that support an official theory
- **Data interpretation** — they suggest that faced with data that conflict with their existing theories scientists adopt different strategies to avoid changing them, including ignoring, rejecting, excluding and reinterpreting.
- **Memory** — information related to an individual's theory will be easier to recall
- **Communication** – only information that supports a given theory will be communicated to like-minded scientists

Brewer and Lambert also argue that in the presence of strong bottom-up information – for example a very obvious and unexpected fact, the role of top-down theory driven influences are diminished [26].

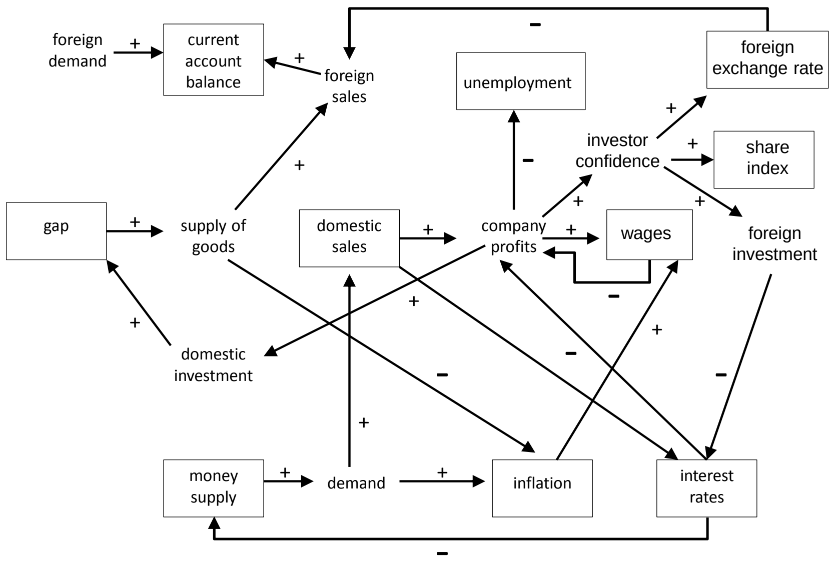

Figure 2.3: Representing a variable based domain theory — from [40]

2.5.3 Theory representation and induction

From the EBM perspective, Donoho *et al.* review a number of constructive theory induction systems, which have been developed to manipulate and revise domain theories, including MIRO, EITHER and KBANN [48]. However, many of these use theory representations which are based on the relationship between variables; for example Clark and Matwin [40] use the representation depicted in Figure 2.3 to define a model of the economy in terms of the set of concepts and entities it contains and the sign of the relationships between them and other concepts — for example an increase in *wages* has a negative relationship with the level of *company profits*. This format is appropriate in an equation based model where structure of the model is expressed in terms of the relationships between a set of variables, and there is no possibility of an emergent variable. In PcGets, theory is introduced by the initial specification of the general unrestricted model (GUM) which incorporates all of the variables believed to be possible candidates based on a combination of theory and previous evidence. It is assumed that the GUM incorporates the real data generating process, and that the variables that are not part of it are deleted from the specification [34]. In equation based variable selection systems like PcGets and RETINA [161], adding or removing variables is equivalent to adding or removing elements in the theory.

2.5.4 Duhem-Quine under-determination hypothesis

Duhem-Quine says that no test is definitive since no theory is tested in isolation. An empirical test is conducted in the presence of auxillary hypotheses about things including constants, empirical evidence, and testing techniques. If contradictory evidence is found it can be unclear whether the problem is with the theory itself or the auxillary hypotheses [88].

Goldfarb *et al.* consider theories to be empirically equivalent when they have identical empirical consequences, and that since a trivial change to a theory will create a new theory but the empirical consequences will remain the same, there are an indefinite number of rival theories that may be created. They identify two broad categories of under-determination of theory by data (UTD), strong UTD which is highly sceptical of the scientific enterprise and in which choosing between theories based on empirical data is impossible in principle, and weak UTD which maintains that although data in science is often imperfect, if the right data were available, theory choice could sometimes be based on empirical evidence [79].

Goldfarb *et al.* investigate whether rival theories of smoking can be proven using empirical evidence, or whether they are under-determined. They consider smoking to be a fruitful area of analysis since theories that aim to explain smoking behaviour originate from different disciplines including sociology, law and economics. They identify two broad classes of theory: rational-choice theories which hold that smokers make rational, reasoned choices based on preferences and non-rational-choice theories which hold that the smoker has not made a competent decision, due to some kind of temporary or permanent impairment of his decision making. Within the range of rational choice theories they identify differences between them due to variations in their treatment of whether or not preferences remain consistent over time, whether or not an agent is expected to have complete information, and the degree to which an agent's rationality is bounded — even agents with complete information may have limited or erroneous decision making capabilities.

They look at five pieces of empirical evidence to help them determine between the different theories: consumer response to price incentives; expressed regret by smokers; brain imaging evidence; quitting behaviour; and subjective risk assessment. They conclude that the outcome is typical of weak UTD in that no single theory dominates the others, based on the empirical evidence available, but that data may become available which would help to refine the choice.

2.6 Marketing

While the preceding sections have looked at agent based models and theory in general terms, this section focuses specifically on marketing theory and literature relating to existing agent based modelling work that has been done in the marketing domain, setting the scene for the applied work to follow. The marketing domain was chosen because of the increasing relevance of social networks, technology and data, which made it a profitable area for applied exploration.

2.6.1 Marketing Theory

The following factors are considered to be important in shaping the purchase process at a micro, individual consumer level:

- **Demand** — It is assumed that households have a level of utility for a product — which may vary substantially across households, with some manifesting high levels of demand in a particular product category and others exhibiting none. Within each houshold, income changes may also affect demand for a product or the price that they are willing to pay [163].

- **Seasonality** — East [51] points out that very few Easter Eggs are sold at Christmas. Demand for a product may vary systematically over time due to exogenous factors like climate, or historical cycles that persist even though the external influences no longer Products follow different seasonal patterns which may increase or decrease aggregate demand [51]

- **Pack-size** — It is often considered that consumers will have lower utility for additional product beyond their level of demand [163]. Based on their level of demand , their storage potential and the product's perishability, consumers may prefer a certain pack size to others.

- **Purchase frequency** — Purchase frequency emerges from a combination of usage rate and pack size.

- **Branding** — Shaw and Merrick [183] cite a recent taste test in which, when the brands being tasted were unnamed, 51 percent of people preferred the taste of Pepsi, but when named, the proportion fell to 23 percent. By making a brand distinct, Kotler and Keller [118] argue that a consumer can organise their thoughts about a product, potentially reducing risk and saving time. Kotler and Keller [118] cites the American Marketing Association's definition of a brand a name, term, sign, symbol, or design, or a combination of them, intended to identify the goods or services of one seller or group of sellers and to differentiate them from those of competitors. In some senses it is similar to a tag in a one-shot transaction — in the sense that it implies a level of trust despite the absence of personal knowledge of the individual seller. Shaw and Merrick [183] note that most consumers are loyal to more than one brand.

- **Repeat Purchase Loyalty** — Numerous studies look at whether past purchases play a role in determining future purchases, either due to memory, habit, loyalty or successful trialling. Bass *et al.* [19] investigate whether or not the *zero order hypothesis* of independence between consecutive purchases holds in empirical data, finding that the result varies between categories. Habit also saves time [51]. In some senses this is similar to the notion of norms.

- **Variety Seeking** — This is the opposite of loyalty and implies a degree of *satiation* with a particular product or its attributes, and is likely to lead to switching. Like repeat purchase loyalty, this also implies a feedback mechanism from previous purchase events [19].

- **Product attributes** — Products have a range of tangible features, which may appeal to particular consumers needs or tastes. Roberts and Lattin [169] point out that a consumers beliefs about a product are more important than the actual characteristics of the product.

- **Price** — A commonly cited theory is that consumers evaluate a product's price by comparing it to a *reference price*. This may be internal – for example a price remembered from the last shop, or external — for example a *regular retail price* or a competitor's price. Prices expressed in a certain way have proven in studies to be be more attractive to consumers, e.g. a 9.99 tag vs a 10 tag [118]. Roberts and Lattin [171] identify a nonlinear response to price in the Australian telecoms market. Below a certain price, consumers may be suspicious of a product's quality [183].

- **Interaction** — One of the most significant social developments in the last decade has been the rise of online social networking sites such as Facebook, Twitter and Linkedin. These sites have helped to draw attention to a phenomenon that has probably always been an important factor in marketing success the role of social pressure and word of mouth recommendation. Social pressure and product diffusion have been explored by Bass and others, and numerous studies have looked at the importance of personal recommendation. These social phenomena are perpetuated by peer to peer contact, and the resulting pattern of diffusion or recommendation can be said to emerge from the interactions between individuals. Emergence in this context refers to a higher order structure that emerges from a lower order system, due to local interactions between the individual components of the system. Because they emerge from interactions between individuals, social media effects can be more unpredictable than the response patterns of many broadcast campaigns. Although seeding activities like promotion and advertising can be important in introducing the message into the network, a message that is widely shared does not necessarily follow the established decay patterns of a broadcast message. Some influences are considered to spread as a result of consumer to consumer interaction, including positive or negative word of mouth. Also, conformity to a majority behaviour — based on a threshold for the percentage of neighbours adopting a product, can be an influence on individual behaviour [47]. Blogging and online reviews are increasingly significant in transmitting information betwen consumers [201].

- **Advertising** — Shaw and Merrick [183] suggest that most advertising is unprofitable. Even so, many studies show that it can have a significant impact on consumer perception, especially when used to communicate other product features such as price and offers [51].

- **Offers and Promotions** — Promotions can have a major short term impact on sales [183], driven through a variety of mechanisms including price reduction, mulitbuys, extra frees or giveaways. Promotions can have a knock-on impact on the timing of purchases, product consumption rates, delayed repurchase due to pantry stocking.

- **Distribution** — If a product is not stocked at the retailer the consumer has visited, it cannot be considered or chosen. Except for a few core, widely stocked products, the choice of retailer pre-determines which products enter the pool of candidates. Shaw and Merrick [183] argue that distribution is one of the most important contributors to sales.

- **Framing effects** — The context in which a choice is presented can affect a consumer's utility for a particular product. Studies have shown that relevant context includes the other choices available, and the default option proposed. Zhang and Zhang [212] cite a study in which participants were asked to choose between a five-star restaurant, which was a 25 minute drive away and a three-star restaurant, which was 5 minutes away. When a decoy choice was added — a four-star restaurant 35-minutes away, participants tended to prefer the five-star restaurant which was 25 minutes away. When the decoy was a two-star restaurant, 15 minutes away, the participants' preferences switched to the three-star restaurant 5 minutes away. Kahneman [105] cites the example of insurance offered to drivers in Pennsylvania and New Jersey. In Pennsylvania the default is a fuller policy, with 79 percent take up, whilst in New Jersey the default is a constrained policy, with take-up of the fuller policy at only 30 percent.

2.6.2 Marketing Instruments and Decisions

Anderson argues that companies have goals and objectives, and create functional plans to achieve these with the aim of maximising their return on investment [7]. He suggests that corporate strategy is the link between the goals and objectives the organisation wants to achieve and the functional plans it uses in its day to day activities. One of the functional policies which they employ is marketing, a social and managerial process by which they obtain what they need and want through creating, offering, and exchanging products of value with consumers [117]. The nature of the marketing policies they might deploy has changed and adapted over time to conform with the requirements of the relevant era [148]. Companies deploy these policies in a market, made up of a group of consumers who are potential buyers of a product or service [43].

Marketing managers are required to make strategic decisions regarding these policies on an ongoing basis, often in complex and dynamic commercial environments [130]. A decision involves identifying possible courses of action and choosing between them, and can involve many attributes, including amongst others, changes to product features, price and advertising deployment. These and other instruments are often called the *marketing mix* — variables which a marketing manager can use to influence a brand's sales in the market [193]. Some of these instruments are described in Section 2.6.1. The impact of different instruments in the mix may be variable, and each brand may have a unique sensitivity to consumer response to its marketing action [144]. In order to effectively maximise their return from their marketing investment, companies need to have knowledge of the processes which generate market-share figures and, be able to analyse the impact of their own actions on market shares [43] to formulate the levels or combination of these variables to maximise sales, market share, or profit [193].

2.6.3 Models to Support Marketing Decision Making

One way of identifying the impact of previous policy decisions and their associated actions is to look at how sales responded to these actions historically[193]. Empirical analysis of marketing processes often involves an encounter between an existing theory about the market and the performance of different policies [184], and new data which may or may not be consistent with the theory. In a simple modelling scenario the existing hypotheses can be relatively limited. For example, an analytical process that involves comparing the effectiveness of online display creative executions operates in a relatively simple theoretical world with few entities and rules to consider, needing only to reason about different types of advertising execution, the number of advertising impressions delivered, and their possible impact on a web user. However the broader marketing environment can often contain many other factors, including pricing changes, promotions, and distribution, with response to one factor conditional on the level of others [131]. An equation-based modelling approach, sometimes referred to as *marketing mix modelling* [193] is often used in commercial marketing contexts to achieve this, identifying the impact of different types of marketing instrument, such as price, promotion and distribution. Tellis gives an example of :

$$log(Y_t) = a + B_1 log(A_t) + B_2 log(P_t) + B_3 log(R_t) + B_4 log(Q_t) + e_t \tag{2.1}$$

where *Y* represents Sales, *A* is Advertising, *P* is Price, *R* is Sales Promotion and *Q* is product quality at time *t*.

The functional form of the model implies that the dependent variable Y_t is affected by interaction of instruments (A_t, P_t, R_t and Q_t) in the marketing mix [193]. The equation is a linearised version of the multiplicative Cobb Douglas form explained in Section 2.3.2.

Practitioners aim to separate the contributions of such instruments to sales or other metrics of interest, and calculate the impact and ultimately the profitability of using them. Corporate policy makers are then able to use these calculations to revise their plans in order to help them achieve their goals. Analysis is typically carried out using a combination of historical time series data and some type of statistical modelling approach, often involving a form of regression under the assumption that the results allow practitioners to predict how consumers might respond in the future and therefore how best to plan their marketing actions [193].

2.6.3.1 Automating Marketing Decision Making

Several existing pieces of research point to a move towards automation of the modelling and decision support process. Traditionally, making sense of complex environments has been associated with the use of expert judgement [54] rather than automated data-driven analysis systems. These have been criticised for their potential to produce outcomes that are theoretically counter-intuitive or unusable in the target domain [189]. A number of trends are driving an increase in the use of automated marketing analytics systems to aid decision making, including: the requirement for faster reaction times to maintain competitive advantage — especially in online marketing environments; the increasing volumes of real time data available to fuel decisions; an increasing focus on return on investment from marketing activities; a proliferation of digital media channels; and shorter product life cycles [23].

In an EBM context the computational issues resulting from testing the numerous models that arise from many combinations of variables, along with the desire to maintain certain characteristics in the selected candidates, have fuelled the development of automatic model selection processes [94]. These are said to outperform manual selection in terms of formulation, variable selection, unbiasedness of estimation, and their ability to consider all relevant evaluation criteria. Some of the automatic approaches used are simple stepwise routines, in which variables are sequentially added or deleted from an equation based on specific criteria [34]. Others, like PcGets and RETINA, are more developed approaches which aim to create models with particular statistical and theoretical properties [161]. The problem of path dependence is addressed in some EBM approaches, like PcGets, which performs multipath searches, working back from every possible deletion from the initial GUM [34], but remains in others like the stepwise procedure. The RETINA procedure acknowledges the impossibility of exhausting all paths and uses correlations with the dependent variable as a guide in exploring a limited number of higher potential possibilities [161]. Genetic algorithms have been used successfully for model selection in mass spectrometry [204, 27], offering the ability to search areas of the solution space which would be closed to hill-climbing methods.

Automated analytical methods have been developed for a range of applications and are a widely used approach in physical [204, 27] and social sciences [94]. Examples of applications in marketing include: the Promoter system — designed to automatically evaluate the impact of promotions [1]; the use of tracking codes in online advertising to identify chains of consumer behaviour and suggest improvements to the allocation of marketing investment [182]; online display creative optimisation [16]; keyword bid optimisation in search engine marketing [25]; and automated product recommendations, based on analysis of the history of a user's characteristics and past purchase behaviour [130].

2.6.4 Agent Based Modelling Applications in Marketing

There is an increasing recognition that social interaction is a key influence in marketing decisions that needs to be accounted for in marketing decision models. Many existing approaches have been criticised for their tendency to treat individuals as isolated individual consumption units and ignoring the influence of customer communication networks [213]. Some of the more significant social influences include:

- **Online social networking sites** — a significant development in the marketing landscape in the last decade has been the rise of online social networking sites such as Facebook, Twitter and Linkedin. Numerous studies have shown how user content can spread virally on social sites like Facebook and Twitter [210, 190, 160].

- **Web 2.0** — the impact of social media extends beyond social networking sites into the mainstream web through consumer review sites like Tripadvisor and collaborative projects like Wikipedia [106]. Beyond specific sites, social media has a significant impact on online behaviour; including search, conversation, community formation, tagging, and content creation and sharing [112].

- **Social advertising** — advertisers are using social interactions to target advertising based on the relationship of an individual to other users or entities [213]. Social advertising is a growing area which makes use of some interaction from a connected user in the target users' network — with the resulting advertisement including reference to that user's interaction. For example, an advertisement for a charity may mention that a user's connection has recently become a fan of the charity.

2.6.4.1 Simulating Patterns of Online Social Interaction

In an EBM context, social pressure and product diffusion have been explored by Bass [19] and others, and numerous studies have looked at the importance of personal recommendation. These social phenomena are perpetuated by peer to peer contact, and the resulting pattern of diffusion or recommendation can be said to emerge from the interactions between individuals. An example of an aggregate equation that takes account of social interaction is below:

$$M_{t+1} = (sg - s)M_t^2 + (1 + sL - f - g - sgL)M_t + gL + aX_t \tag{2.2}$$

where M_t is the number of individuals with the target property/ total individuals in population, L is the upper limit of individuals who are susceptible to the property, s gives the gain or loss from social interaction, g represents fixed gains from a constant source, f represents the probability of members who have the property leaving the group, X_t is an exogenous influence, and a a parameter.

Huckfeldt *et al.* [97] developed this equation using a first order differential structure in which the main elements of a social diffusion process are reflected. The equation above aims to reflect a social process in which the total level of subscribers to a particular belief, M at time period $_{t+1}$ is made up of losses, which are subtracted, and gains which are added, due to social interaction and other factors.

There are a number of recent studies of how information propagates across online social networks, covering Twitter, Facebook, Flickr and others. Each network has a slightly different information sharing mechanism.

Sun *et al.* [190] use data supplied by Facebook to examine how information diffuses amongst its users. They find that on Facebook it is rarely the case that a small number of nodes creates a chain reaction. Instead, information is likely to enter the system via many independent sources and as a result large clusters can emerge when these short diffusion chains merge together. Facebook diffusion chains can be extremely large (up to 82 levels have been observed) and are generally a result of a multiple chain-reactions. They suggest that the length of these chains may be partly due to the sheer number of individuals on Facebook as well as the ease of performing actions on the site.

One of the main information propagation mechanisms on Facebook is the News Feed feature. It differs from some other mechanisms in that it does not require specific action on the part of the user. Recent actions are broadcast by default to the user's network of friends. Celebrities and artists, ideas and interests, businesses, etc can represent themselves on the network and interact with users via the Pages product. Users can become a fan of the page, post messages, etc. These actions are then broadcast to their friends' News Feeds.

Page diffusion occurs when a user becomes a fan of the page. This is then broadcast on their friends News Feeds. Friends observe this action and decide whether or not to also become a fan. The chain-starters characteristics have not been found to be a significant factor in determining the length of the chain [190].

Yu and Fei [210] have access to data, collected by crawling the Flickr network, that contains the *Favourite* photos for each user, along with the date and time that the photo was marked as a favourite. In total the dataset contains almost 35 million Favourite marking events against 11 million distinct photos The *Favourite Photos* feature allows users to keep a publicly visible set of photos, which are shown to their contacts upon login. The photos are arranged such that more recent additions to a favourites list are more likely to be visible than older ones.

Because a relatively small group of users account for a large proportion of transactions on the site, the authors focus on a subset of the data that covers usage for a single, connected cluster of users, accounting for 25 percent of the total user group. The data contains transactions over 100 days, with no record of the state of the network before or after the 100 day period.

The study considers that an *infection* has taken place if a user marks as a favourite a photo that has already been marked by one of their contacts. If a photo has been marked by multiple contacts it is given multiple attributions. Because of this, when cascades of transmission are identified in the data, some chains of propagation overlap with other chains.

The authors create a simple generative model of the infection process. A node (Flickr user) is selected at random and its state is set to infected. The node's contacts are infected with probability beta. This beta is assumed to be constant for all photos, so does not take account of different levels of interest in different photos. The original node is set to uninfected, so can be infected multiple times. The newly infected neighbours infect their own neighbours. The progress of the resulting cascade is recorded. The authors calibrate the beta parameter to 0.035 equivalent to a 3.5 percent probability that transmission will occur as a result of any single contact event. They then run 5 iterations of the model and compare the resulting artificial cascades with actual cascades found in the data. They find that the size and distribution of the chains compares well with chains within the real data. They experiment with different seeding approaches, and also with decaying the probability of transmission as the photo moves away from its initiating node, but find that the simpler model provides a better fit [210].

Cha *et al.* [35] test the theory of how influentials affect networks by studying the structure and manifestations of influence within the Twitter network. They use a dataset of 2 billion links between 54 million users who between them, in the period analysed, made 1.7 billion tweets. The authors gathered the data by crawling 80 million userids in August 2009. They found that 95 percent of users were connected in a single giant cluster, with the remainder either having no connections or being part of a separate network. The largest cluster accounted for 90 percent of tweeting activity. They identify three types of influence that can be observed amongst Twitter users:

- **Indegree influence** – the number of followers a user has — there is some discussion about whether indegree necessarily implies influence on Twitter, since users have a tendency to reciprocate follows (links).
- **Retweet influence** – the number of times a user is retweeted retweeting may result in propagation of a tweet into the wider network, beyond one-to-one interaction with a user's own connections.
- **Mention influence** – the number of mentions containing a user's name usually representing a public response to a user's tweet.

The study compares how individual users score on each measure of influence. They find that users who are frequently retweeted are also likely to be frequently mentioned, but that indegree is not correlated to either of the two other scores in other words users with more followers are not more likely to have their messages spread.

The study also looks at whether users who score highly in terms of their likelihood to be retweeted have a similar influence across all of the topics that they mention. To do this they reviewed tweets about three specific events that had generated interest on Twitter: the Iranian presidential election, H1N1 and the death of Michael Jackson. They find that the level of influence demonstrated is similar across different topics.

2.6.4.2 Innovation Networks

Delre *et al.* [47] seek to address the question of why more than 50 percent of products introduced into the market are failures. To do this they build and trial an agent based model to simulate the adoption of new products by consumers. They simulate a range of factors that might affect the launch of a product — splitting them into two broad categories. They identify internal factors, such as transmission of information by word of mouth, and external factors like mass media and promotions. They also look at how testing the timing and targeting of the different factors can contribute to the development of strategic insights to improve the success rate of product launches.

The agents in their model are arranged on a *small world* network that falls between a completely regular structure and a completely random one. Information enables the agents to enter the decision process if at least one neighbour has adopted the product, or if the agent has received a mass-media message. The agent has a specific preference for the product in question. In addition, the agent has a threshold for social pressure, based on the percentage of adopters in his immediate network so that once a number of agents around it are behaving in a particular way then it will begin to behave like them. Each agent's utility threshold for the product is based on this combination of product preference and social pressure.

The simulation starts with seeding a sub-group of the population is switched to be users reflecting the introduction of a compelling external promotion. The word of mouth process then spreads outwards from the seeds. If the process stops the company have to organise a new promotion. They can also use mass media, which has the effect of allowing an agent to be part of the decision with a given probability. There are three possible states that a consumer might be in: **Non-aware**, **Aware and non-adopter** or **Aware and adopter**. This structure makes it possible to identify two social influence effects one is social pressure, the other peer to peer product information transfer an agent can be aware but not adopt.

The authors run a series of tests on this hypothetical structure, varying the promotional targeting approach, and the timing of promotions and mass media. They hold the other parameters in the model constant. The S shaped curve is common in marketing take-up literature. It reflects the balance of external and internal influences on product take-up. The authors look at two possible promotional targeting strategies: *throwing rocks* targeting a cohesive group of highly connected consumers with a promotion increasing contagion within local area, and throwing gravel randomly assigning the product to individuals maximising spread across groups but less chance of achieving social pressure within any one group.

Running a series of simulations, they find throwing gravel to be inefficient. They suggest that the best strategy is to use a combination of strategies. Using the model, the authors attempt to identify the optimum time to run mass-media to maximise the seeding and WOM impact, for different strengths of campaign. The authors report that a typical approach is to position seeds and then launch a mass-media campaign to support the seeding. Because of the contemporaneity of the two effects, in reality it's often unclear which is having the bigger impact. They find that it is only worth starting a mass-media campaign if there is 10 percent take-up of the product, because otherwise there will be insufficient support from social influence. They hypothesise that in the absence of any social pressure, consumers made aware of the product will decide not to buy it.

Rand and Rust [166] present a set of guidelines for using ABM (Agent-based Modelling) and illustrate them with a model of consumer adoption of innovations, based on the original Bass model. They replicate the original Bass results on the adoption of innovation and then incorporate the existence of a social network among the agents. In the illustrative model agents have properties p (probability of adopting due to mass media) and q (probability of adopting due to word-of-mouth effects). The authors suggest using Agent-based models as computational experiments, varying different model inputs to see how outputs are affected (they stress that when ABM incorporates stochastic elements multiple versions of each scenario need to be run and then an average taken). Applying this approach to the adoption model they find that in networks where consumers do not know many other consumers and the relationships between consumers are distributed more unequally adoption increases more slowly than in networks where there is a moderately high level of connections.

2.6.4.3 Consumer Simulations using Multi-agent Systems

There are a number of existing studies which are relevant to the marketing simulations explored in later chapters.For example Schwaiger and Stahmer[177] develop a simulation model which aims to be a realistic representation of a single store or supermarket, including goods and customers. The model has several distinct groups of agents including a supermarket agent containing customer agents and product agents which provide information about the store, such as the goods on offer, different prices and promotions, the layout of the shelves, sales and cost data and others. The customer agents are composed of a personal profile which contains information on the customer's age, income, gender and domicile. Empirical data on customer characteristics is extracted from customer cards data, questionnaires and general marketing knowledge and encoded in a set of conditional probability rules, which are stored in Bayesian networks to represent their interdependencies. The data is then used to model customer behavioural characteristics.

Lavington [123] creates an individual level model that simulates the weekly buying behaviour of a representative consumer panel. It is designed to be suitable for the market of inexpensive, fast moving consumer goods. Customers are divided into the geographic areas they live in and have demographic characteristics which are used to determine their behavioural characteristics. Individual characteristics such as brand attitudes, product usage and reading habits can differ among customers in the same customer categories. In the simulation companies have four strategies available to them — distribution, price, promotions, and advertising. For retailers, these strategies determine what percentage of retailers stock the brand and the display prominence it receives. For customers there is a two-stage effect. First, a customer's conditioning, which determines the preference for the different brands before entering the shop, is governed by price, promotion, advertising, past usage of the product, and personal recommendations from acquaintances. Second, once the customer enters the shop, they are also influenced by factors such as on-pack offers and display so the final total purchase disposition is a combination of the effects in the two stages.

The total purchase disposition is determined by customer conditioning and size preferences. Conditioning depends on three factors – price, advertising and usage. The price effect depends on the relative prices of the products and any coupons received by the consumer. The advertising effect is a result of the brand's advertising expenditure since the last time the customer went shopping and its level depends on the intensity of impressions produced and the level to which they affect the consumer. The usage effect is calculated depending of the consumer's past usage of each brand (including any free samples) and its magnitude depends on the consumer's current brand attitudes. Size preference effect is also taken into account to arrive at the total repurchase disposition the consumer has a certain initial brand and size preference ranking but these can be affected by different in-store offers, which is accounted by a switching matrix which is applied to the initial preferences depending on the current in-store offers. Once the final purchasing probabilities for each customer are calculated, random sampling is used to determine the predicted level of purchases by customer.

North *et al.* [154] describe a virtual environment that represents the shopping behaviour of consumer households and the business behaviour of retailers and manufacturers in a simulated consumer market. Consumers in the model are represented by consumer households which are comprised of shoppers, inventories and users. A household has one or more shoppers, and inventory to store purchased items and users who consume the items. Shoppers choose where to shop depending on their individual store preference which can evolve over time depending on store experiences. They choose items to purchase depending on their consideration sets (out-of-store (OSCS), and in-store (ISCS)). The OSCS is the set of products considered before a shopper enters a store and are determined by weighted preferences of the brands on offer

Retail stores are simulated as locations to purchase goods and have shelves with products, advertising flyers, product displays price reductions and others associated with them. Retail stores are also grouped in neighbourhoods which contain a set of competing stores and shoppers. They also contain a representative of each retail channel (e.g. both individual, drug and club stores) which take up a different proportions of total stores, determined by user input.

The authors stress that their model does not aim to predict future behaviour of market participants or market outcomes but rather to test the robustness of various marketing strategies and discover potential trend drivers by generating ranges of potential market outcomes for different strategies.

2.7 Conclusion

In Chapter 1 the primary objective for this thesis was defined — the development and evaluation of a solution to the problem of selecting an agent based model specification and its associated theory from the possible candidate specifications — and in Section 1.1.1 the overall solution to the problem was decomposed into five constituent research requirements. This chapter has presented a review of the literature that is relevant to these research areas, and collectively these sections have provided a review of the literature in fields relevant to automated model specification in ABM and introduced the domain of marketing, and in particular, marketing models and decision making. It also drew attention to two key trends in the domain: an increasing need for automation and a developing recognition of the importance of social interaction in marketing processes.

In the sections below, the literature reviewed is considered in terms of its implications for the five research areas entailed by the overall problem, laid out in Section 1.2. In summary these requirements were for an agent-centric theory representation, a method for mapping a theory to a model, a technique for scoring candidate models, an algorithm for searching the candidate space, and a method for interpreting the outputs of the models with their associated theories.

1. **An agent-centric theory representation** — From the theory representation viewpoint, most existing work — like that by Clark and Matwin [40] — represents theories in terms of concepts and the relationships between them. This format is appropriate in an equation based model where structure of the model is expressed in terms of the relationships between a set of variables, and there is no possibility of an emergent element. However in ABM a representation of a theory specification needs to be able to accommodate the possibility of emergent factors developing, such as peer to peer communication. Similarly, existing theory representations used in automated EBM systems, for example PcGETS [94] and RETINA [161], are limited in their ability to represent interaction and emergence, since they relate to models in which these elements are not directly specified.I extend

the existing work in EBM by developing a theory representation that is agent-centric — in other words that is based on the characteristics and behaviours of individual agents rather than aggregated concepts.

The literature reviewed in Section 2.3.8 is helpful in defining the additional requirements of an agent-centric theory representation that can represent interaction and emergence. In particular Conte's [42] conceptualisation of different types of emergence and downward causation provides clarification of the different processes that could be represented. Muller's [151] emphasis on the role of observation, and the particular perspective from which the emergent process is observed by an external observer is informative in designing the framework of a theory structure in which the emergent fact may have an impact on agents regardless of whether or not the agent is being directly influenced by other agents through interaction. I build on these ideas by incorporating them into a computational representation that is part of a learning framework.

The ideas put forward by Delre *et al.* [47] and North *et al.* [154], discussed in Section 2.6.4.2 regarding modelling of social diffusion are a good basis for the design of a social diffusion structure. Watts and Strogatz' small world network design methodology is a suitable platform for creating simulated networks [202]. In addition, the practical applications of Cha *et al.* [35], Sun *et al.* [190] and Yu and Fei [210] — particularly applying aggregate level models to an online social diffusion process and linking the drivers of diffusion to specific elements of online content — contain many concepts and ideas that inform the design of the social transmission mechanisms in later chapters.

In terms of defining the boundaries of a theory representation, Epstein's conceptualisation of a *base*, modelled ontology, which is influenced by a *total* ontology [56] is illuminating. Existing case studies such as the work on the Artifical Anasazi model [102, 58] are also informative in this regard, in that they provide examples of bounded models with endogenous processes which are also subject to external influences.

2. **A method for mapping a theory to a model** — In the context of my research, the form and structure of the model used needs to be able to reflect a theory, and needs to be parameterisable by the search process, so the concepts that need to be represented in the model are related to the domain to which it is being applied. because of this, some of the concepts that are used in other domains, such as tagging, reputation and norms are less directly relevant to my research, while others, such as communication, social network simulation, and topology are more immediately applicable.

   The level of detail at which the mapping needs to be performed is a significant factor in ABM since many modellers emphasise a bottom-up approach in which the models aim to simulate the underlying mechanics of the data generating mechanism, and existing work on complexity [41, 55, 57, 139] helps to inform this requirement.

3. **A technique for scoring candidate models** — In terms of criteria for choosing between theories, work by Kuhn [119] and Wacker [200] provides some core criteria for evaluating the qualities of rival theories. I build on this by making some of the criteria automatically evaluable in a theory representation. From the point of view of assessing compliance with a theory, there is existing work in EBM, where concordance with a theory may be expressed by the combination of variables being selected in the model, the functional form of the model, and a coefficient having an expected magnitude or a coefficient taking an expected sign. The requirement in an ABM is different in the sense that there are other elements in the model, such as the form of a social network, which need to be taken into account. Therefore I build on this approach by arranging the theory representation in the form of an ontology, meaning that the theory can be assessed for other characteristics.

4. **An algorithm for searching the candidate space** — From the point of view of model selection, existing research in both EBM and ABM forms a starting point for my research. From an EBM perspective, various model selection methods have been developed [31], some of which are based only on fit to empirical data, such as R-Squared, and others which combine elements of theory. The variable selection methods put forward by [93] and others for an EBM environment form a part of my overall approach to model selection which I build on by formulating an ABM approach that allows more model elements to be made available to the selection process. In addition, from the parameter calibration standpoint, work done by Fabretti [61, 60], Terano [194] and others in calibrating agent based models provides a useful platform to extend and customise in the applications developed in later chapters.

5. **A method for interpreting the outputs of the models and their associated theories** — A key requirement for interpreting the outputs of a model is that the relationship between the inputs and the outputs can be understood. One way of achieving this is by variance decomposition. Some work has been done in this area by Saltelli and others [174, 173, 129], exploring the first order and global impact on the outcome for each of the input variables. This work is valuable in helping to conceptualise the way in which a model interpretation algorithm should work.

   As well as the ability to quantify the scale of an input's effect, understanding the probability that the input factor is really having that effect is also important in interpreting a model. Although it is out of the scope of this research, work done in applying the statistical theory of maximum likelihood to agent based simulation [115] and its applications, particularly in computational finance [136, 4, 127], is useful in conceptualising the role of uncertainty in an automated interpretation system.

As well as these specific areas, the literature review informs the requirements of the solution to the overall research problem, laid out in Chapters 4 and 5. In the next chapter, Chapter 3, many of these ideas are applied in a simulated competition between two modelling approaches, ABM and EBM, intended to determine the value of an ABM approach in an applied context. Then in the following two chapters the existing research discussed in this chapter is extended and built on as the automated modelling design is developed and evaluated.

Chapter 3

Evaluating the Performance of Automatic Model Selection Methods Using Agent and Equation Based Models

3.1 Introduction

The chapter builds on and synthesises elements of the previous two chapters. It describes a simplified application of the automated theory selection approach outlined in Chapter 1, and then evaluates its performance in a simulated environment which involves many of the features and methods described in Chapter 2. The primary aim of the chapter is to evaluate whether an automated agent based methodology, which provides an initial solution to the problem of selecting agent based models, has an advantage over existing methodologies.

To achieve this a simulated experiment is carried out to compare the effectiveness of an automated agent based model selection method with that of an automated EBM approach. The performance of the automated model selection processes is evaluated in a series of encounters between corporations in a simulated market, with the rival corporations using contending approaches to evaluate their decisions and guide their policy choices.

Section 3.2 gives a high-level overview of the simulation structure and its characteristics. Then in Sections 3.3 through to 3.5 there follows a detailed account of each agent's characteristics. This is divided into a description of their goals, characteristics, behavioural rules and possible actions. The execution of the simulation is then detailed in Figure 3.4 before the results of the simulation are presented in Section 3.6. Finally in Section 3.7 the results are considered in the context of the overall research objective.

3.2 Simulation Overview

This section gives an overview of the simulation, and positions it in the context of some of the key areas reviewed in Chapter 2. As discussed in Section 2.6.1, social interaction is increasingly recognised as

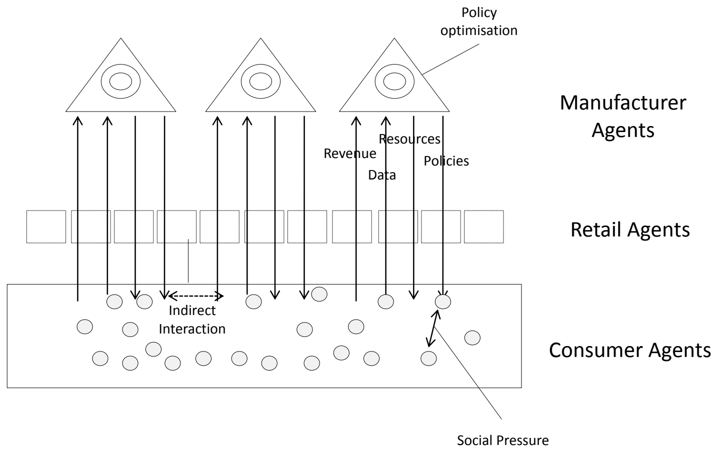

Figure 3.1: Consumer market simulation overview

an important factor in the marketing domain, and as highlighted in Section 2.6.3.1, a number of factors are driving an increased need for automation in marketing decision making. Bringing together these two themes, a simulation study was conducted to compare the efficiency of EBM and ABM models, created through an automated learning approach, under different levels of interaction in the target process.

To achieve this a simulated consumer market was created containing three types of agent:

1. **Consumer agents** — These agents represent customers in the market. The manufacturer agents, described below, seek to learn the consumer agents' preferences in order to maximise their sales. Apart from the policies deployed by the manufacturers, social pressure also plays a role in the consumer agent's decisions. The consumer agents interact with each other through a network. The aggregate buying behaviour of the consumers represents the total demand for the goods being exchanged and the collective output of the manufacturers represents total supply in the market.

2. **Retailer agents** — Retailer agents represent shops in the market, they mediate the exchange between manufacturers and consumers, and implement instructions passed from the manufacturers.

3. **Manufacturer agents** — These agents represent the companies which supply the nominal product — each of the manufacturer agents aims to optimise its marketing policies to achieve the greatest share of purchases from the consumer agents who form the virtual market. The market nominally represents a fast moving consumer goods market, but it could equally represent another market with similar characteristics. The interactions between the manufacturer agents are mediated by the consumer

agents, for example if a particular manufacturer attracts consumers with its policies in a particular period, those consumers are not available to the other manufacturer agents in that period.

The success of three competing companies, manufacturers *A, B* and *C* are compared in a series of simulated multi-period encounters, in which the manufacturer agents use one of three different methods to learn the impact of their policies — either an agent based model, an equation based model, or a random policy selection. The agents exist in a non-spatial environment, and the simulation is enacted over a number of iterations that represent weeks.

Figure 3.1 shows the interactions between the different types of agent in the simulation. The manufacturer agents interact with their environment, made up of the consumers in the simulated market and, indirectly, with the competing manufacturers. In other words, the interactions between the manufacturers are mediated by the purchasing decisions of the consumers. In addition the consumer agents interact through a social network structure, described in Section 3.5 below.

3.3 Manufacturer Agents

Having established an overview of the simulation in the previous section, this section outlines the characteristics of the manufacturer agents. In Section 3.3.1, the different learning approaches used by the manufacturing agents are described, then in Section 3.3.2 their respective model calibration methods are laid out. In Section 3.3.3 their possible actions are defined before finally in Section 3.3.4 their action selection rules are detailed.

3.3.1 The Manufacturer Learning Process

As discussed in Section 2.6.2, a common factor in theories of corporate goal setting is the maximisation of return on investment [7]. In this simulation, each manufacturer agent aims to maximise its return in terms of sales through its choice of policies to use in the market. Choosing the best policy is therefore key to achieving their goal of maximising sales.

In the simulation, manufacturer agents *A* and *B* select their policies based on their theory about how consumers respond to their policies in the market. In order to associate the impact of its policy deployment with the state of the environment, and measure the magnitude of the different policy effects it is able to use, it needs to create a modelled representation of the environment which can reflect the consumer decision making processes.

Each manufacturer has information about the policy (at) which it has deployed in each time period t, and the state of the environment at time t in terms of the number of sales it has achieved in the market. However, it receives no direct feedback about the consequences of it's policies so must infer information indirectly. There are three types of policy selection approach used in the simulation, with two of them involving a modelling process and the third random selection. The selection approaches are described below:

- **Manufacturer *A* — Equation Based approach** — Manufacturer *A* uses an equation based approach to learn the impact of its policies in the market. The agent uses the hypothesis that the data generating process can be expressed using an aggregate representation of the model's inputs and outputs, but with an unknown set of variables and unknown coefficients. It uses the following equation as its representation :

$$S_A = int_A + \sum_{i=1}^{n} w_i * P_{Ai} \tag{3.1}$$

where S_A represents the total sales for manufacturer *A*, int_A represents an estimate of a base level of sales, w_i represent the policy impact coefficients to be estimated and P_i represent the candidate policies that may form part of the selected model.

The equation above reflects manufacturer *A*'s viewpoint that by calculating weightings, w_i, on the input variables for his policies P_{Ai}, it can predict its sales S_A at an aggregate level for a particular point in time. In other words, by calculating a weighted sum of the input policies it aims to explain the level of sales that it achieves in a particular period.

- **Manufacturer *B* — Agent Based Model approach**— Manufacturer *B* hypothesises that the market should be represented as an agent based model, with unknown variables and coefficients at work. The manufacturer creates a representation of the market using a system of agents with the ability to respond to the specific policies of manufacturer *B*. In manufacturer *B*'s model the consumer agents calculate their utilities for manufacturer *B*'s policies as follows:

$$UB_j = wbp + wsp * SPB_j + \sum_{i=1}^{n} wpb_i * PB_i \tag{3.2}$$

where UB_j is the utility of agent *j* for manufacturer *B*, wbp is an overall parameter estimated across all of the agents, wsp is an overall parameter estimated across all agents, SPB_j represents the social pressure experienced by agent *j* to adopt manufacturer *B*, wpb_i represents the overall preference level — for each of the i policies of manufacturer *B*.

The equation above shows the agent's additive utility function, in which each of the attributes of the manufacturers policy (PB_i), and the social pressure (SPB_j) they experience are multiplied by the preference weights and added to the base level of utility (wbp) to give an overall degree of preference for the post (UB_j). The agents in manufacturer *B*'s model are assumed to experience different levels of social pressure, but are homogenous in their in their level of base preference and in their preferences for the manufacturer's policies — only one weight (wpb) is estimated across all of the agents for each of the manufacturer's i policies.

Each agent compares their utility for manufacturer *B* against a threshold value, so that total sales for manufacturer *B* in each time period are the sum of agents for whom utility for manufacturer *B* exceeds their threshold value:

$$S_B = \sum_{j=1}^{k} (UB_j > TB_j) \tag{3.3}$$

where S_B is total sales for manufacturer *B*, *k* is the total number of agents in the model, UB_j is each agent's utility for manufacturer *B* and TB_j id each agent's utility threshold.

In other words, the total predicted level of sales achieved by manufacturer *B*, S_B is equal to the sum of the agents for whom utility exceeds their threshold value.

The utility function used in manufacturer *B*'s model, described above, differs to the data generating model outlined in Section 3.5 in that manufacturer *B* is only concerned with understanding whether agents are buying his own product, and because he has no knowledge about the policies of the other manufacturers. In addition manufacturer *B* makes the simplifying assumption that the agent's preferences are homogeneous, whereas in the data generating model they have individual preferences for the different manufacturers and for different policies. Apart from these differences the consumer agents in manufacturer *B*'s agent based model have the same characteristics as the consumer agents described in Section 3.5.

- **Manufacturer *C* — Random approach** — manufacturer *C* chooses its strategies at random with equal probability.

3.3.2 Model Calibration

This section describes the calibration processes used by the two manufacturer agents (*A* and *B*) who are conducting modelling exercises.

3.3.2.1 Existing Theory

As discussed in Section 1.1, selecting the correct model for a particular domain may need to take into account a combination of empirical evidence and existing background knowledge about the field. Both manufacturers *A* and *B* have existing core beliefs about the domain, represented in their pre-conceptions about the role of the different factors in the market, which in turn constrain the models which they are willing to accept. Their shared existing theory is that:

- *The effect of advertising on consumer utility should be >= 0.*
- *The effect of promotions on consumer utility should be >= 0.*
- *The role of social pressure on consumer utility could be positive or negative.*
- *A consumer's base preference for a manufacturer could be positive or negative.*

These prior beliefs act as a restriction on the range of theories that will be considered. The method by which these restrictions are applied is detailed in Figures 5.3.2.1 and 3.3.

3.3.2.2 Model Search and Evaluation

Manufacturer agents *A* and *B* read in data about their sales in previous periods, along with a record of the policies that they have deployed in each period — both datasets are indexed by time period.

Both manufacturers *A* and *B* use this equation to assess the fit of their models:

$$ModelFit = 1 - \left(\frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2} \right) \tag{3.4}$$

where y_t is the value of the target variable and $\hat{y}_t$ is the modelled value for a particular time period.

The model fit calculation above is a measure of the degree to which the predicted value of aggregate sales, $\hat{y}_t$ recreates the real level of aggregate sales y_t over time. A higher level of model fit implies that the model has a greater ability to re-create the true level of sales.

The processes that manufacturers *A* and *B* use to select their models, taking into account both prior beliefs (listed in Section 3.3.2.1, and empirical evidence (expressed as $ModelFit$ above), are detailed below:

Both manufacturers *A* and *B* use a genetic algorithm to search for the best policy to use. As Michaelwicz [142] observes, any abstract task to be accomplished can be thought of as solving a problem, which, in turn, can be thought of as a search through a space of potential solutions, and since the best solution is sought, we can view this task as an optimization process. Genetic algorithms are an approach that carries out a multi-directional search by maintain ing a population of potential solutions which evolves across generations. In each new generation the good solutions reproduce and the bad solutions die. Each solution is represented by a chromosome, and some members of the population are subject to the operations of crossover and mutation. Crossover involves combining the characteristics of two parent chromosomes to form offspring by exchanging some of their features. Mutation changes some of the characteristics of a particular chromosome by randomly changing some of the features with a probability equal to the mutation rate. Mutation has the effect of adding extra variability into the population [142].

- For manufacturer *A*, an automated variable selection process is used. This means searching for a combination of variables in an equation such that when calibrated the signs of the coefficients associated with the variables do not violate its existing beliefs. The selection process is laid out in Figure 5.3.2.1. For each model in the population of models estimated in a particular iteration of the algorithm, the ordinary least squares solution [84] to the equation is calculated. Equations with combinations of variables that lead to sign violations are scored poorly, whereas combinations that lead to signs that are consistent with existing beliefs are scored highly. The process iterates until the best model is found that is consistent with existing beliefs.

- Figure 3.3 details the workings of the process used by manufacturer *B*. Although, like manufacturer

Figure 3.2: Manufacturer *A* equation selection process

1. Create initial population of 150 equations with bits in each chromosome representing variables absent or present in the equation and initial values selected at random with probability of 50%. In other words each variable has a 50% chance of being selected as present in each equation.
2. Repeat until termination criteria met
   - (a) For each equation in the population
     - i. Estimate model coefficients int_A and w_{ij} in equation 3.1 using ordinary least squares [84]
   - (b) Score models using fit (using Equation 3.4) and theory compliance criteria defined in Section 3.3.2. Models in which the signs of the w_{ij} for the respective policies correspond with the statements about those policies set out in Section 3.3.2.1 are scored highly.
   - (c) Create new population of equations. Reproduce equations proportionally to score
   - (d) Random crossover on top 20 equation pairs with two parents and probability 50%
   - (e) Random mutation on top 30 of the new equation population with probability 5% in each bit
   - (f) Check for termination criteria — whether all w_{ij} comply with theory compliance criteria in Section 3.3.2 and 100 additional iterations completed
   - (g) Loop, if not terminating
3. Exit procedure if terminating

A, the approach also uses a genetic algorithm, rather than searching for a combination of variables it is searching for parameter values for the coefficients associated with the policy variables. The parameter range that the algorithm is able to explore is constrained before the search begins by the manufacturer's beliefs. For example if a prior belief specifies that the parameter must be an integer between 0 and 3, the possible values submitted to be explored by the algorithm would be $[0, 1, 2, 3]$.

3.3.3 Possible Actions - Manufacturers

Having outlined the process through which the manufacturers assess the impact of their policies above, this section explains the possible policies that the manufacturers are able to deploy. Each manufacturer can deploy only one policy in each time period. A manufacturer agent may:

Figure 3.3: Manufacturer *B* theory selection process

1. Create initial population of 150 agent based models with each bit in the chromosome representing a parameter value for the parameters wbp, wsp, and the wpb_i parameters associated with the PB_i in equation 3.2. The initial parameter values drawn at random from permissible (theory compliant) parameter range defined by the criteria defined in Section 3.3.2.
2. Repeat until termination criteria met
   (a) For each agent based model
      i. Evaluate fit of the model using Equation 3.4
   (b) Score agent based models based on level of fit
   (c) Create new population of models. Reproduce models proportionally to score
   (d) Random crossover on top 20 equation pairs with two parents and probability 50%
   (e) Random mutation on top 30 of the new equation population with probability 5% in each bit
   (f) Check for termination criteria — whether iterations exceed 200
   (g) Loop, if not terminating
3. Exit procedure if terminating

1. **Advertise** — As mentioned in Section 2.6.1, studies [183, 51] suggest that advertising can have an impact on consumer perception. In the simulation, manufacturer advertising is visible to the consumer agents, and increases the utility for that manufacturer amongst those with a preference for advertising.

2. **Promote using either price or coupons** — As discussed in Section 2.6.1, promotions can have a major short term impact on sales [183], driven through a variety of mechanisms including price reduction, multi-buys, extra-frees or giveaways. In the simulation, manufacturer promotions are visible to the consumer agents, and increase the utility for that manufacturer amongst those with a preference for promotions.

3. **Do nothing** — In this case the manufacturer doesn't deploy any policy.

Each policy has a reward associated with it, although the actual reward may differ to that which a manufacturing agent believes it has, depending on the accuracy of the theory it is using. The level of the reward is reflected in the consumer agent's utility for the particular policy, detailed in Section 3.5.

3.3.4 Action Selection - Manufacturers

The manufacturers select which action to take using one of two selection strategies.

- If the simulation iteration number $<= 20$, each manufacturer chooses a policy at random
- If the simulation iteration number $>= 21$ then:
  - Manufacturers *A* and *B* choose the policy which, based on a review of the estimated policy weights learnt from their learning exercises (described in Section 3.3.2) in iteration 21, maximises their expected level of sales.
  - Manufacturer *C* continues to select its policies at random.

3.4 Retailer Agents

The retailer agent in this simulation is passive — it exists to mediate change in distribution policy. Since the manufacturer is not the end seller, it stocks product or not and deploys feature or display on the instructions of the manufacturer. It also varies price at the instruction of the manufacturer.

3.5 Consumer Agents

The consumer agents in the model have the following characteristics:

- **Goals** — In the simulation, the consumer agent's desired outcome is to acquire the product for which it has maximum utility.
- **Social Network** — A range of existing research, reviewed in Section 2.6.4, looks at the simulation of social networks. In this simulation, the consumer agents are connected by a network which allows the

transmission of social pressure based on Delre *et al*'s [47] social pressure model. The pressure felt is based on the percentage of adopters in his immediate network so as the proportion of neighbouring consumer agents increases so does the level of pressure to conform:

$$SP_{ij} = \frac{N_j}{N} \tag{3.5}$$

where SP_{ij} is the social pressure felt by consumer agent *i* to purchase manufacturer *j*'s product, N_j is the number of consumer agents with whom consumer agent *i* is linked who last purchased manufacturer *j*'s product, and *N* is the total number of consumer agents to whom consumer agent *i* is linked.

In other words, as the proportion of the agents who are linked to agent i and are already purchasing the product increases, the social pressure on *i* also increases. This social pressure is fed into the utility model shown below.

A small world network was created using the methodology proposed by Watts and Strogatz [202] — assigning each of the consumer agents to a node on a regular network, then randomly rewired the links until the network showed the properties of a small world — a mixture of short paths connecting most of the individuals within each clique, and longer paths connecting the cliques.

- **Utility Function** — The consumer agents calculate their utilities for each manufacturer's policies as follows:

$$U_{mj} = bp_{mj} + SPB_{mj} + \sum_{i=1}^{n} * pref_{ij} P_{mi} \tag{3.6}$$

where U_{mj} is the utility of agent *j* for manufacturer *m*, bp_{mj} represents agent j's base level of preference for manufacturer *m*, SPB_j represents the social pressure experienced by agent *j* to adopt manufacturer *m*, $pref_{ij}$ represents the individual preference of agent j for policy i, and P_{mi} represents the value for each of the i policies of manufacturer *m*.

The equation above shows the agent *j*'s additive utility function, in which each of the attributes of the manufacturers policy (PB_i), their preference for each policy ($pref_{ij}$), their base preference (bp_{mj}), their social pressure preference ($sp_p ref_j$) and the social pressure (SPB_{mj}) they experience are cross-multiplied and summed to give an overall degree of preference for each agent j for each manufacturer m — U_{mj}. The agents in the model are heterogeneous in their base level of preference for each manufacturer, their level of preference for social pressure and their preference for each of the manufacturers policies. In addition they experience different levels of social pressure. The agent's individual preferences are described below:

  - **Advertising preference** — Consumer agents are given a level of preference for advertising drawn from a random uniform distribution between 0 and 1. The average preference for advertising across all of the consumer agents is therefore 0.5.

- **Promotions preference** — Consumer agents are given a level of preference for promotions drawn from a random uniform distribution between 0 and 0.5. The average preference for promotions across all of the consumer agents is therefore 0.25. On average the consumer agents prefer advertising to promotions.

- **Base preference** — Consumer agents are given a level of preference for each manufacturer drawn from a random uniform distribution between 0 and 1. On average the consumer agents have a base preference of 0.5 for each manufacturer, meaning that in the absence of other policy deployments each manufacturer would be able to sell equal volumes of their product.

- **Purchase Cycle** — This is drawn from a random uniform distribution between 0 and 5, consumer agents purchase when they have reached the end of their specific cycle. The cycle increments in each simulation iteration, and then resets to one when the purchase cycle number is reached, so for example an agent with a purchase cycle of 4 will purchase every 4 iterations.

3.6 Simulation Results

Figure 3.4 gives an overview of the dynamics of the simulation in each iteration. In this section the results of the simulation are presented and reviewed. Each encounter between the manufacturer agents lasts for 52 periods, representing one year. The level of social interaction was implemented at 6 different levels, creating a total of 600 simulated encounters. The simulation was run 100 times for each level of social interaction in the consumer market.

Figure 3.5 shows an example of a single run of the simulation. It shows the levels of sales achieved from the consumer agents whilst the manufacturers randomly deploy their policies during the first twenty iterations of the simulation. After they have completed their modelling exercise at period 20 the manufacturers then select what they believe to be their optimal policy to deploy in the remaining periods. In this run, manufacturer *B* has successfully learnt the optimal policy, while manufacturer *A* has made an incorrect assessment and manufacturer *C* has randomly selected a poorly performing choice.

Figure 3.8 shows the policy selection successes for different manufacturers in the 600 runs of the simulation. At low levels of social pressure (between 0 and 50%) the performance of manufacturers *A* (using EBM) and *B* (using ABM) in discovering the optimal policy to use was very similar, with a success rate of around 90 to 100%, while manufacturer *C*, using the random strategy, deployed the correct policy on roughly 33% of occasions. As the role of social pressure in the market was increased, moving towards 100% of the performance of manufacturer *A* (using the equation based model) began to reduce towards the level of a random strategy, while manufacturer *B* using the agent based model approach continued to perform well in selecting the correct policy. At 100% social pressure, the EBM approach was able to determine the

Figure 3.4: The possible actions that each agent may take in the simulation

- In each time period each consumer agent may:
  1. Evaluate whether it needs re-supply, based on its purchase cycle
  2. Commence product evaluation if it needs re-supply
  3. Calculate its utility for each of the possible choices
  4. Purchase the product for which it has the highest utility
- In each time period each manufacturer agent may:
  - *If iteration <= 20*
    1. Randomly choose one of the possible policies (from range of policies shown in Section 3.3.3)
    2. Act — deploy its selected policy
  - *If iteration = 21*
    * Manufacturer *A* updates its model of the consumer's theory using the algorithm shown in Figure 5.3.2.1 and the previous 20 periods as a training dataset, then chooses what it believes to be its optimal policy.
    * Manufacturer *B* updates its model of the consumer's theory using the algorithm shown in Figure 3.3 and the previous 20 periods as a training dataset, then chooses what it believes to be its optimal policy.
    * Manufacturer C chooses a policy at random
  - *If iteration >= 21*
    1. Act — deploy the optimal selected policy chosen in the model update process (from range of policies shown in Section 3.3.3)

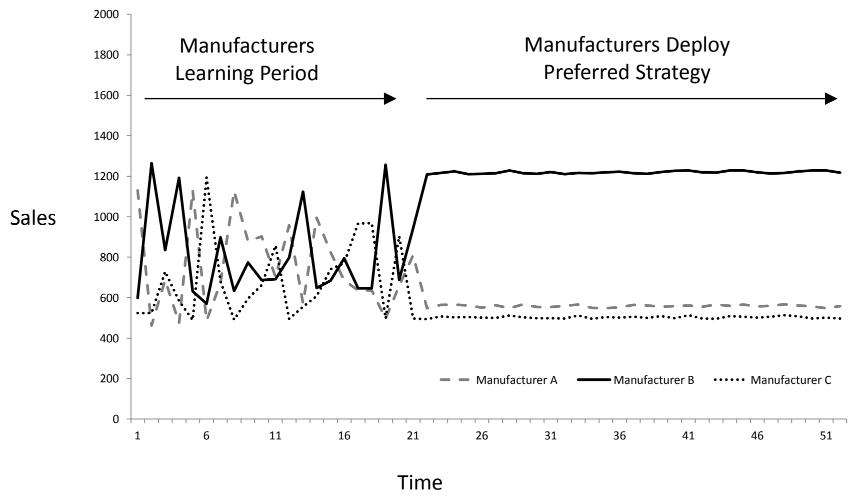

Figure 3.5: A single run of the learning simulation

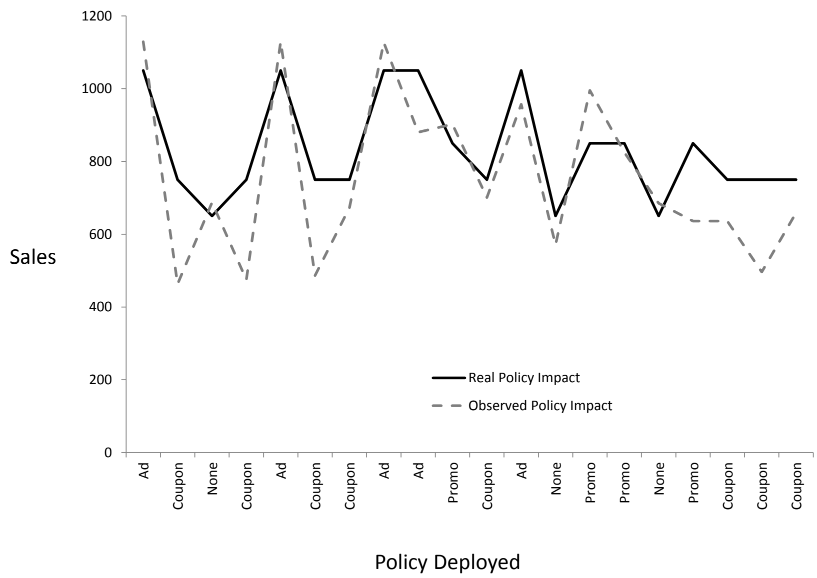

Figure 3.6: Noise introduced by the policies of competitor manufacturer agents

correct policy to use on only 33% of occasions - the same level of success as that of the random selection strategy.

The learning achieved by manufacturers *A* and *B* is inherently disrupted by the noise in the market introduced by the policies of competitors since competitor manufacturers are simultaneously taking actions which may counteract the impact of the manufacturer's policy, meaning that it is unable to achieve an accurate reading of its impact. Figure 3.6 illustrates this, showing an example of one time series view for a particular manufacturer agent. The figure shows the difference between the actual impact of his own policies and the time series of sales that is realised. The difference between the two lines is due to the confounding activities of the behaviours of the other manufacturer agents and social pressure. The policy deployments of the competitors are not directly observable to the other manufacturers — only indirectly through changes in the state of the environment.

For manufacturer *A* the level of unexplained variation in sales is greater. Since social pressure is not one of the hypotheses he is able to test using the EBM approach, there are movements in his own sales which are not due to either his own or competitor policies, but rather the endogenous effect in the market of the consumer agents influencing each other.

Figure 3.7 shows the rising standard deviation in sales as the level of social pressure increases — in other words the sales levels achieved become more volatile as the level of social pressure increases. These results are most easily understood in the context of the utility function (described in Section 3.5) used by the consumer agents. Since the agents have a base level of preference for each manufacturer, in the absence of any other policy interventions their product selection would be the manufacturer for whom they have the maximum base preference. As the level of social pressure applied in the simulation increases, sales patterns become more volatile because the effect of base preference, which helps to keep the levels of sales across the three manufacturers stable, is overwhelmed by the impact of social pressure between the agents.

3.7 Conclusions

This chapter presented a simplified version of the automated model selection approach proposed in Section 1.1.1 — aimed at solving the problem of selecting agent based models and their associated theories. In Section 3.2 an overview of the simulation was presented, followed by a detailed description of its workings in Sections 3.3 through to 3.5. The dynamics of the simulation's execution were explained in Figure 3.4, and the simulation's results were presented in Section 3.6.

The conclusions that can be drawn from this chapter fall into a number of areas:

- *Evaluating the proposed framework* — In terms of the overall research objective outlined in Section 1.1.1 — to develop and evaluate a solution to the problem of selecting an agent based model specification and its implied theory from a group of candidate specifications — the solution appears to be promising in that the agent based approach used by manufacturer *B* performed well in the simulation, proving to be able to select the correct model to maximise its profitability, and outperforming

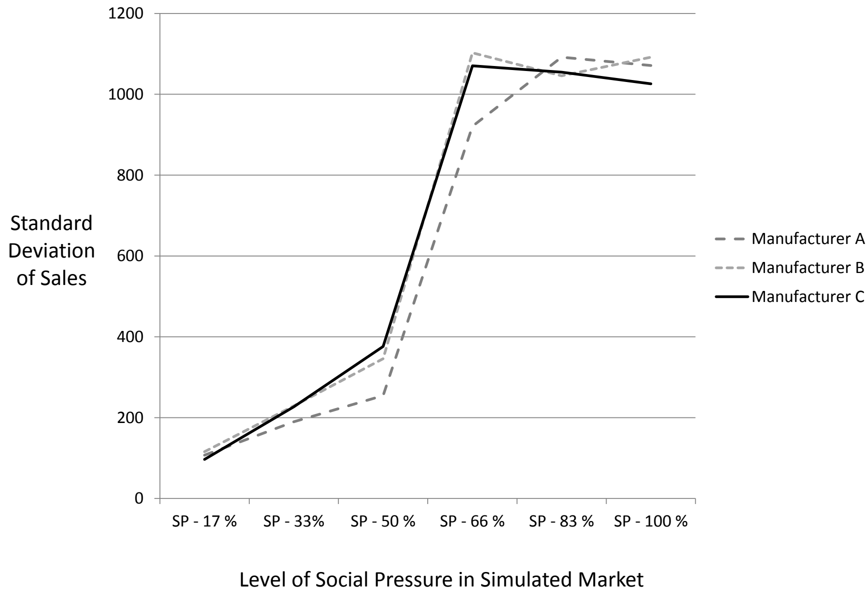

Figure 3.7: The impact of increasing social pressure on the standard deviation of sales in the simulated market

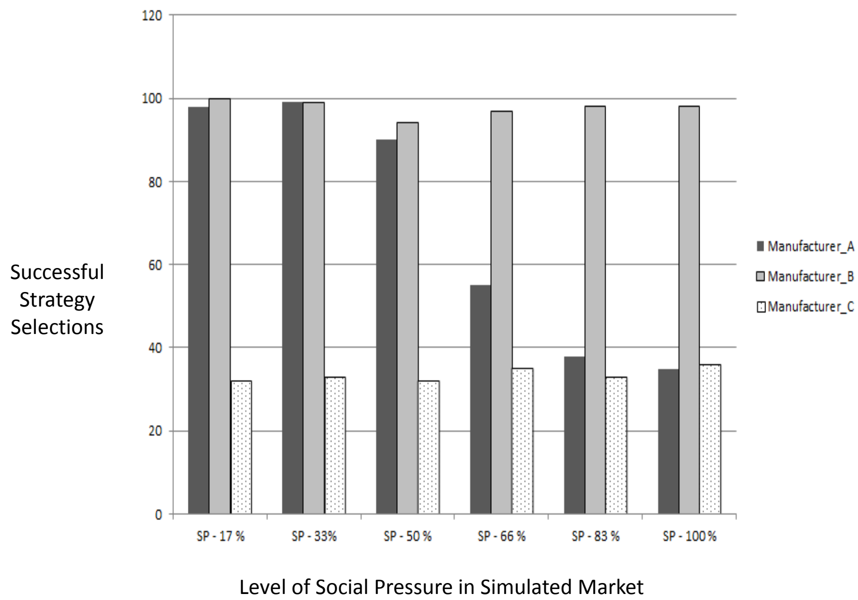

Figure 3.8: The impact of increasing social pressure on optimal policy selection

manufacturer *A* which used the existing EBM methodology.

As discussed in Chapter 1, there are three key themes that underpin the proposed solution — theory, data and models. The automated agent based approach used in this chapter demonstrated several functionalities relevant to these themes. Manufacturer *B* brought its existing theory about the possible impact of its policies (defined in Section 3.3.2) into the analytical process and used them to constrain the hypotheses that could be explored in its modelling exercise. It then successfully maximised the fit of the agent based model to a data set, using the approach described in Figure 3.3, while taking into account restrictions imposed by existing beliefs about the domain. It was also able to simulate emergent behaviour and use the simulation to learn about the level of impact emergent behaviour was having in its environment. Finally, it used the new theory it learnt about its policies through the modelling process to determine its policy selection (laid out in Section 3.3.3) in the simulated environment. As discussed in Section 3.6, variations in the market that could not be explained with the data and models available disrupted the success of the EBM approach in finding the optimal policy choice. This kind of noise could pose a threat to the ABM approach too under different circumstances.

- *Broader methodological issues* — From a broader methodological standpoint, it was clear that at low levels of social interaction a similar, genetic algorithm based automatic modelling approach worked well for both the ABM and EBM model specifications. This suggests that although the forms of these two types of model are different, successful model selection has components that are common to both — efficient search methods and a constrained hypothesis space. As might be expected, the advantage of the ABM specification becomes more apparent where the simulated data generating process in the consumer market is more reliant on interactions between agents. In these cases the internal forces in the process create significant movements in the simulation outcomes that the EBM specification used by manufacturer *A*, focussed on measuring the impact of external effects, struggles to account for. Since the simulation was conducted at a high level of abstraction the implications for corporate marketing strategy are more limited.

- *Applied marketing research* — From the point of view of the two key marketing related themes discussed in Chapter 2, automation and social interaction, the approach demonstrated in this chapter shows that in principle it is possible to create an analytical system that incorporates both capabilities.

Given that the automated ABM approach has shown success in both areas in the simulated competition, the methodology is developed further, with the aim of solving the problem of selecting a model and theory from a set of possible alternatives, in the chapters that follow — Chapters 4 and 5. Then in Chapters 6 and 7 further evaluation of the method is conducted against real empirical data.

Chapter 4

Theory Representation and Mapping to Models

4.1 Introduction

In Chapter 1 the primary objective for this thesis was defined — the development and evaluation of a solution to the problem of selecting an agent based model specification and its associated theory from the possible candidate specifications — and in Section 1.1.1 the overall solution to the problem was decomposed into five constituent research requirements. In summary these requirements were for an agent-centric theory representation, a method for mapping a theory to a model, a technique for scoring candidate models, an algorithm for searching the candidate space, and a method for interpreting the outputs of the models and their associated theories.

This chapter proposes solutions to the first two of those five research requirements, and the chapter that follows, Chapter 5 proposes solutions to the remaining three. The solutions presented in this chapter build on the ideas put forward in Chapter 3, in which a simplified automated model selection approach proved successful in a simulated environment. In each section that follows the relevant features of that simulation are referenced and the main methodological extensions highlighted. The chapter also draws on the literature review in Chapter 2 to define the sub-requirements of each of the solutions proposed. The two areas to which solutions are proposed in this chapter are:

1. *Theory representation* — An agent-centric theory representation is developed that takes into account the emergent characteristics resulting from agent interactions. The necessary characteristics for such a representation are outlined in Section 4.2.1, and then a design that fulfils the criteria is laid out in Section 4.2.2.

2. *Connecting theories and models* — A rule-based mapping protocol is proposed that allows the theory specification to be mapped to different types of ABM to support testing across a variety of methods and applications. In Section 4.3.1 the required features of such a protocol are discussed, then in Section 4.3.2 a mapping system is described that demonstrates these features.

To put these two research requirements in context, Figure 4.1 (described in Section 1.1.1) shows the area to which they apply highlighted and located in the overall structure of the automated theory selection approach.

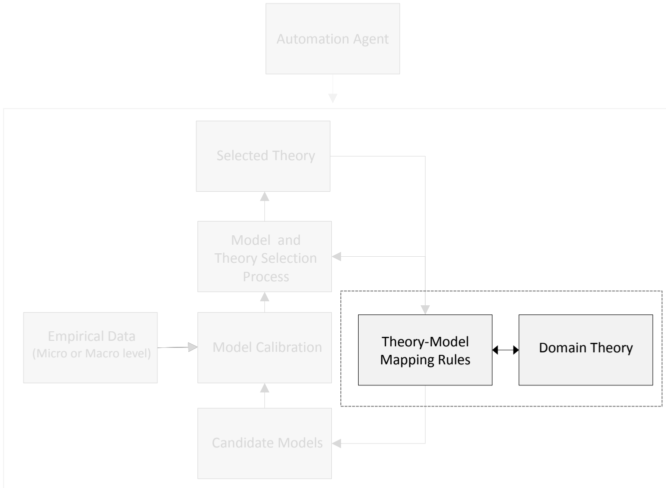

Figure 4.1: The elements of the research design addressed in this chapter

4.2 Theory Representation

This section addresses the first of the five research requirements outlined in Section 1.2 — the requirement for a computational theory representation with which the modelling and data components can interact, which is conformable with the specific characteristics of an agent based model. As initially indicated in Section 1.1.1, the ability to automatically select the appropriate model and theory for a particular domain entails a representation of the theory that can be used to take into account existing theoretical knowledge, and also define the candidate hypotheses that are to be tested against empirical data.

In Chapter 3, theory played an implicit role in shaping the possible outcomes of the models used by manufacturers *A* and *B*, for example in the functional form chosen by the different manufacturers to represent the dynamics in the market, and the variables considered in it. More explicitly, the range of hypotheses that could be searched during the calibration process was constrained by the prior beliefs laid out in Section 3.3.2. All of these constraints had an impact on the outcomes of the models used by the manufacturers, but were applied in an ad-hoc way. The purpose of this section is to provide a generalisable but structured approach to computational theory representation.

4.2.1 Theory Representation Requirements

The theory representation described below in Section 4.2.2 is constructed to take into account a number of requirements that were identified as part of the literature review process documented in Chapter 2. These are:

1. **It should be *bottom up*** — in keeping with the literature reviewed in Section 2.3.8 it is assumed

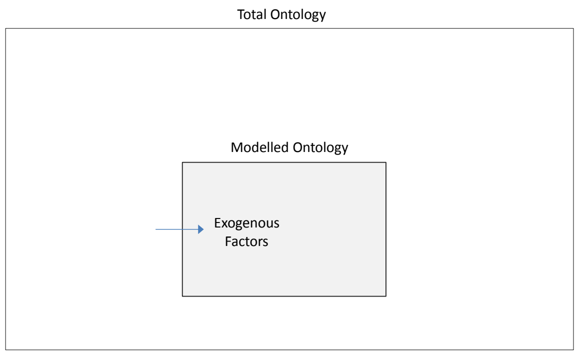

Figure 4.2: Total and modelled ontologies

that an agent based model requires a theory of the causes from which to grow the effect [41]. A theory representation is therefore required which is framed in terms of agents' characteristics and their relationship with the corresponding exogenous and emergent factors which they perceive implying that macro-level behaviour emerges from micro-level behaviour and a chain of causes and effects is mediated through the agents in the system.

2. **It should be interoperable with different model forms** — a wide variety of methodological approaches to agent based models has been proposed, some of which were reviewed in Chapter 2. A successful theory representation therefore needs to be able to interact with different types of agent model, including rule based models and statistical approaches.

3. **It should be able to incorporate information which is determined outside of the model** — since any model is a partial representation of the real world, the theory should be able to incorporate external inputs. It is assumed that there can be facts which are determined outside of the model which have an influence on its internal dynamics, as in the Artifical Anasazi model [102, 58]. In Epstein's conceptualisation, the theory representation is a *base*, modelled ontology, but influenced by a *total*ontology [56]. The theory representation allows features created in the total ontology, which is not represented, to interact with the base ontology. Figure 4.2 shows this in a simple diagrammatic form. The grey area represents the workings of the agent based model, including any endogenous features such as artefacts that emerge from interactions in the model. The total ontology is the white area and includes everything from outside of the model that may have an impact its dynamics.

4. **It should be able to incorporate factors which may emerge during the model's execution** — as discussed in Section 2.3.8, emergence is a common occurrence in AB environments, a theory

representation should therefore be able to represent emergence and downward causation.

(a) *It should be able to represent agent interaction* — the theory representation needs to accommodate an ontological representation of epistemological emergence due to interaction between agents — in other words emergence in which the emergent factor is reducible to the sum of its parts [185].

(b) *It should be able to reflect downwardly causal influences* — in addition the theory structure should be able to represent downward causation, in which the agents mediate causal power from the exogenous and other factors in the model through their impacts on agent's decision making and behaviour, creating an emergent factor through agent interactions, which in turn may have an impact on the behaviour of the agents in the model which is distinct from the impact of the individual interactions [41]. For example, a promotion may cause consumers to discuss a particular product amongst themselves — this is represented by the interaction between agents. However if the global level of interaction becomes known (in other words if an agent becomes aware of the fact that a certain proportion of the other agents are talking about a particular product) this may have an additional impact beyond that felt through his own network.

4.2.2 Theory Representation Design

In this section a structure for the theory representation is proposed that takes into account the requirements identified in the previous section. Using Dzeroski's definition, a theory can be stated using terms from a domain's taxonomy and an interconnecting set of laws [50]. Having evaluated various forms of knowledge representation system, an ontology was chosen as a suitable form of representation for an agent-centric theory since it is able to represent entities of a specific knowledge domain and the relationships that can hold between them [95]. An ontological approach allows the system to represent the elements of a domain's taxonomy, with a set of rules specifying the relationships between the entities [133]. For the implementation required in this research, the ontology identifies the entities involved and the functional relationships between them, with the nature of the relationship between the different elements specified using a set of supporting rules. For example, the ontology element might specify that A and B are objects in the model, and that there is a relationship between the size of A and the size of B. The supporting rule details a specific non-linear function that characterises the relationship between them [132].

Figure 4.3 shows the overall structure of the representation, along with the elements that it contains. These elements are described in detail in the next section.

4.2.2.1 Elements Contained in the Theory Structure

This section details the contents of the theory structure. The structure contains the following elements:

- ***Agent Properties*** — These include general properties of the agent.

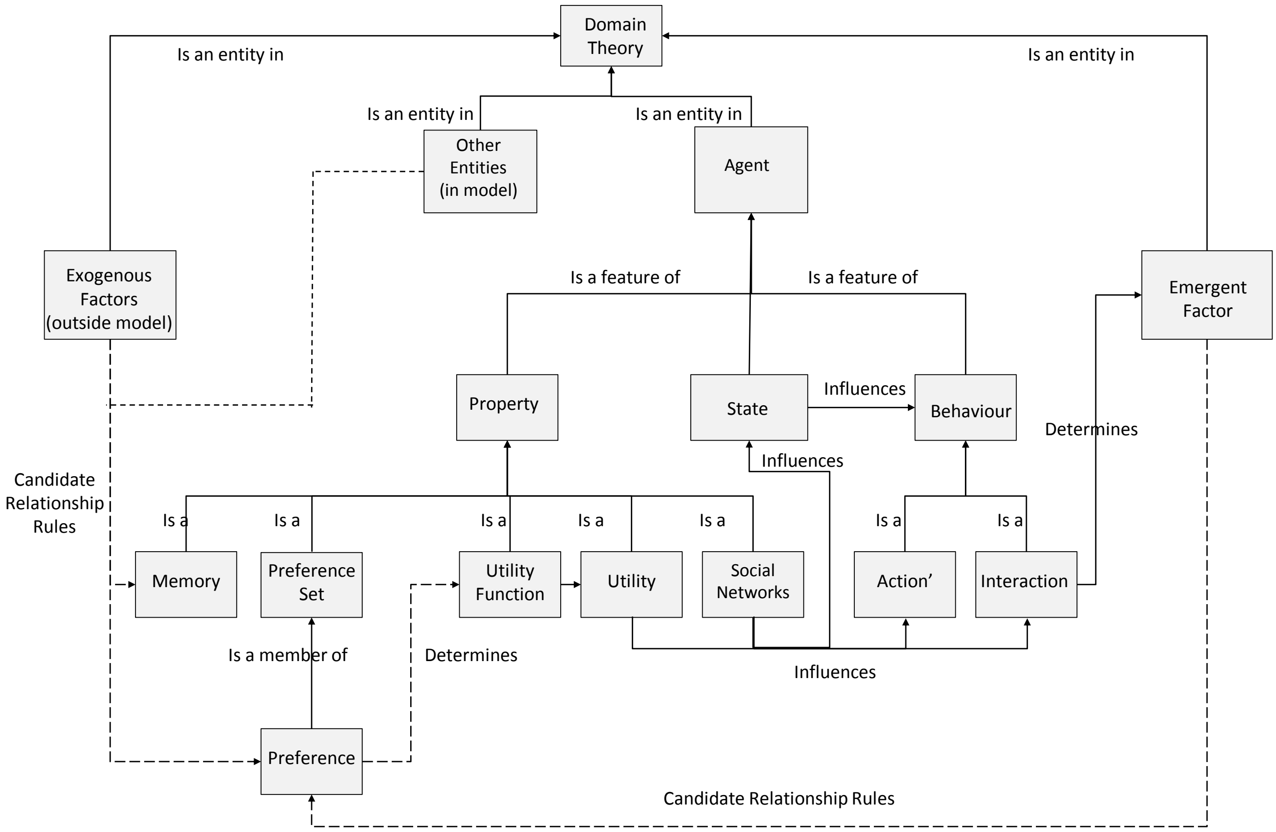

Figure 4.3: The proposed agent based domain theory

- ***Memory*** — Agents can retain some beliefs across time. These may be beliefs about the environment, other agents or their own previous states and preferences.
- ***Preferences*** — An agent in the theory representation has a set of preferences. Preferences are represented as a hierarchy, with higher level preferences linked to sub-preferences. In the scoring process described in Section 5.2.2, the hierarchical preference set structure means that candidate theories can be scored for coherence — the degree to which sets of concepts appear to go together [196].
- **Utility Function** — There are various possible forms that an agent's utility function could take, and multiple functions could be tested as part of the theory search. A very generic utility function that could be used across many modelling applications is a simple additive calculation in which the weights are given to each attribute and the overall sum of the weighted attributes constitutes the agent's utility, like that presented in Equation 3.6.
- ***Utility*** — Each agent may have multiple utilities, affecting for example their decision to share information in a social network, or separately to make a purchase. These utilities are ultimately related through the social diffusion feedback mechanism.
- ***Agent States*** — States can be a useful mechanism for representing agent behaviour, for example for representing enduring changes in an underlying disposition which may change the agent's behaviour until the next state is attained. In these cases agents may be in one of a finite number of states at any particular time. Depending on the context, states can be considered in different ways, for example enduring changes in an underlying disposition which may change the agent's behaviour until the next state is attained. In order to keep the model tractable, the number of possible states that an agent can occupy should be minimised, since as the number of states increases the number of parameters that need to be estimated to model transitions grows massively [29].
- ***Actions*** — An agent's capabilities are the alternative courses of actions it can execute to interact with the world [191]. The range of possible actions that an agent can take need to be defined, based on a combination of observation of the target individual's behaviour and a classification process that establishes which action the target may be performing. The agents have a utility based approach to action selection. The agent makes a choice between possible options based on the expected utility of each, taking into account the state that it is in, its preferences and the attributes of the different options. Each possible choice may have multiple attributes.

• ***Constants*** — These are values that relate to the operations of the model, for example the time-frame under review.

• **Exogenous Factors** — As discussed above, exogenous factors are variables that may have an influence in the model but that whose values are determined separately to its dynamics.

- ***Mappings*** — These relate external data entities to conceptual entities in the model.

- ***Relationship Rules*** — These define the hypothesis space of relationships to be tested, laying out the set of hypotheses about how agent preferences map to exogenous and emergent factors. Each relationship could be represented by a number of possible rules, depending on the strength of the prior belief. For the theory elements that the user has prior beliefs about, statements about the likely effect of that element on a consumer's utility can be defined, taking into account the range of the original variable. By using a normalised utility function, with both the input variables and weights scaled to fall between 0 and 1, a qualitative theory statement such as *has a very strong effect on* can be directly translated into a corresponding numeric utility weight range, for example between 0.8 and 1. These statements are fed into the theory specification.

- ***Emergent Factors*** — These are represented separately to exogenous factors, since they are in part a consequence as well as cause of an agent's utility and so do not map directly from external data. The apparently recursive nature of the representation — in that the emergent effects are caused by and also cause preferences — is avoided through the role of sequencing. The sequencing of agent behaviours within any iteration means that the behaviour of previous agents influences later agents, but that the agent does not influence its own behaviour within a given iteration. The effect of social pressure may also be felt across time through the role of memory.

  The emergent entity is also a dynamic element in the ontology, since it is not known to definitely exist *a priori* and may or may not become a causal factor in the course of the model's execution. Whether or not it does emerge depends explicitly on the causal structure (exogenous factors, rules and preferences) elsewhere in the ontology which generates the agent's micro-level behaviour. Epstein anticipates situations in which the entire causal ontology may need to be dynamic, for example if the model run creates a situation in which a completely unexpected causal factor develops. In this case the ontology exists as a series of different states over the run of the model [98]. In this kind of application the structure of the model is sufficiently limited in terms of factors that the agent can perceive that this situation is unlikely to arise.

  Figure 4.4 shows the flow of causality from the initiating factors through to the impact of downward causation. The figure shows how an exogenous causal factor for which agents have some preference may affect the behaviour of an agent, then through communication or interaction with other agents also affect their behaviour. The emerging collective behaviour — the fact that a number of agents are now behaving in a certain way — may affect the behaviour of agents for whom a trend can be a behavioural influence, regardless of whether or not they interact with agents who are exhibiting the behaviour.

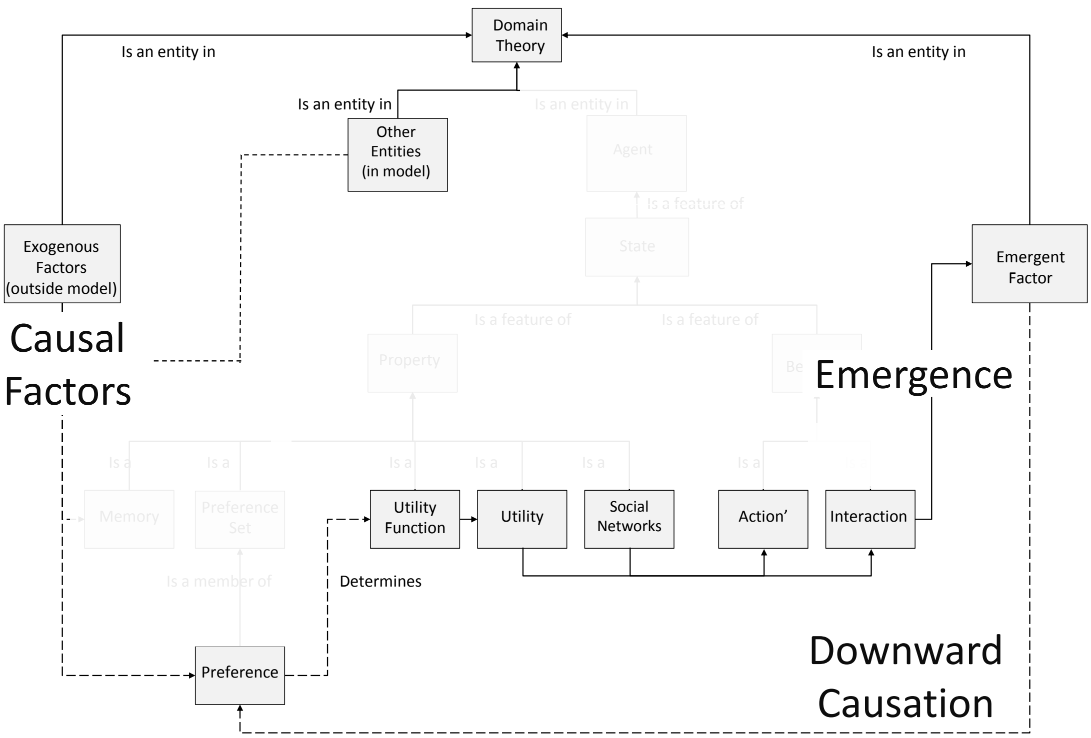

Figure 4.4: Causal factors leading to emergence and downward causation

This section considered the necessary elements of a theory representation that could interact with the other parts of the automated modelling selection system, defined in the next section and Chapter 5.

4.3 Connecting a Theory Representation to a Model

In the previous section a theory representation was defined that could be used to host the theory elements involved in an agent based model selection process. This section is concerned with a method for connecting the theory to the model. This addresses the second of the five research requirements outlined in Section 1.2 — the need for clear communication protocols between the theory representation and an agent based model.

In Chapter 3, there was a defined relationship between the theory and the model used by manufacturer *B*, but it was not mediated through a formalised process. This was partly because the agent had no formal representation of his theory, but also because there was no defined format for it to relate a theory to a model. For example, a range of hypotheses that could be searched during the calibration process was laid out in Section 3.3.2, but the implementation of the constraints required customised adjustments to the searchable parameter range. The objective in this section is to propose a more formal method through which the constraints imposed by a theory, or the range of acceptable theoretical possibilities, can be communicated to the agent based model which reflects the theory.

4.3.1 Requirements for Mapping Rules

In this section the requirements of mapping rules in this context are outlined. The requirements are based on the review of the literature in Chapter 2. The requirements are:

1. **It should be able to connect model elements with theory elements** — to test a theory systematically using a computational model, the relevant components of the model need to be manipulable in line with the elements of the theory. The level of detail at which theory can be embedded in agent based models was highlighted in Section 2.5.1 — in particular because of the *generative* approach favoured by many ABM practitioners. AB modellers often argue that because of its generative, bottom-up approach the theory is built in to the structure of the model, in the sense that the models aim to simulate the underlying mechanics of the data generating mechanism [41, 55, 57, 139]. Figure 4.5 shows an example of a set of mapping rules that would be necessary to relate the elements of a theory specification to their associated model elements. The figure illustrates the connection of the theory elements, via the mapping rules, to the model elements. There may be one-to-many relationships between theory elements and computational model component counterparts — one theory element may map to multiple programmatic elements in the model.

2. **It should be sufficiently flexible to work with multiple ABM approaches** — although a consistent theory specification is assumed, the model to which the theory needs to be mapped might take a wide variety of forms, depending on the application. There is a range of complexity in the behavioural representations used in typical ABMs, from *environmental determinism* at one extreme, in which a

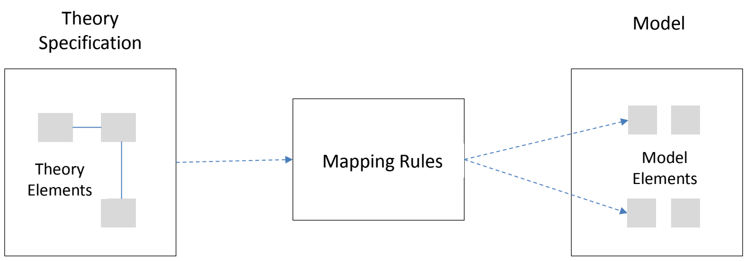

Figure 4.5: Relating theory elements to model elements using mapping rules

limited number of mutually exclusive possible situations each corresponds to an action-mapping, to more complex systems which sustain multiple goals, with actions and sub-actions available to support them. In general simulations with a larger number of agents tend to use more abstract representations, while those with less tend to use more complex representations [29]. Examples of these diverse representations from existing research include:

(a) **Game Theory** — As discussed in Section 2.3.7.2, game theory is frequently used in agent based modelling to represent the possible outcomes from the interactions of agents. For example Axelrod [10] used the Prisoner's Dilemma to model the potential gains from different approaches to cooperation.

(b) **Cognitive Architectures** — As mentioned in a Section 2.2.2, a variety of cognitive architectures have been developed that aim to represent intelligent reasoning systems including ACT-R [38], SOAR [121], and PRS [206].

These requirements form the basis for the mapping file design outlined in the next section.

4.3.2 Mapping File Design

In the previous section, various requirements were laid out that a theory to model mapping protocol would need to be able to accomodate. In this section an approach is outlined which meets those requirements. In the proposed approach, the theory is passed to the model with a pre-defined specification that relates each theory attribute and associated rule to a set of model elements, which in combination implement that part of the hypothesis in the model. The following is an example of a simple protocol for transferring information about the contents of a theory, in which the characteristics of the theory are represented in the sub-sections of the overall rule:

([Preference], [Relationship Rule], [Relevant Utility Function], [active model elements], [active rules], [active parameters], [possible parameter values])

Because the candidate rules defining the possible relationships between preferences and utilities are mutually exclusive, statements could also be passed to the model in terms of AND/ OR groups for example:

(RuleA OR RuleB OR RuleC) AND (RuleD OR RuleE or RuleF) etc..

For example, using the structure above, a hypothesis that social diffusion has a strong impact on agent behaviour can be mapped from the theory representation to a model made up of agents representing consumers might be expressed as:

([Social Pressure], [High Sensitivity], [Buying-utility], [network-on, social-perception-on], [evaluate-active-neighbour-count], [small-world-network], [.5,.7]])

As well as different network types, different interaction processes can also be implemented. For example: the epidemiological/ infection mode is used by Yu and Fei in a study of social cascades on Flickr. Under this model an agent can be in a number of finite states, for example: *Susceptible (S)*, *Infected (I)*, or *Recovered (R)*. To test two possible theories, in which the impact of social pressure on agents is still believed to be high but the mechanism by which social pressure spreads is unknown, the rule could be expressed as:

([Social Pressure], [High Sensitivity], [Buying-utility],[network-on, social-perception-on], [evaluate-active-neighbour-count], [small-world-network], [.5,.7]) OR ([Social Pressure], [High Sensitivity], [Buying-utility],[network-on, social-perception-on], [SIR], [small-world-network], [.5,.7])

As part of this system, the elements of the rules are interpreted and implemented in the ABM.

In an ABM with heterogeneous agents there may not be a single theoretical model which can explain all possible outcomes — for example one agent may be price sensitive while another is not. Heterogeneity can be maintained in the model through a number of possible means, depending on the structure of the target model. A distribution of preferences could be defined amongst the agents in the model itself, so that even though a single central parameter is passed to it, heterogeneous behaviour is maintained because of the heterogeneous utilities that would result. Alternatively, a range of possible values could be passed to the model, such that agents could take on heterogeneous parameters within a set of bounds. An extension of this would be to pass a central mean parameter to the model and also specify a distribution function for the values of the individual agent characteristics. This approach is widely used in Hierarchical Bayesian models [2].

4.4 Conclusion

This chapter put forward solutions to two of the five research requirements outlined in Section 1.2, which collectively form the proposed solution to the research problem of selecting an agent based model specification from a set of contending candidates.

1. The first of these was the requirement for a computational theory representation with which the modelling and data components can interact, and which can conform to the specific characteristics of an agent based model. To address this, the specific features of a theory representation that would be needed to meet this requirement were reviewed in Section 4.2.1. Then in Section 4.2.2 the structure and characteristics of an ontology-based theory representation that met those needs were described.

2. The second research requirement that this chapter addressed concerned the need to be able to relate a theory to a model by developing clear communication protocols between the theory representation and model that specify how a set of hypotheses can be passed to a model. In Section 4.3.1 the characteristics of a protocol that would meet the needs defined in Section 1.2 were identified, then in Section 4.3.2 a mapping protocol was described which would allow the communication of agent-centric theories between a model and a theory representation.

In the following chapter the other three research requirements outlined in Section 1.2 are addressed. These are: a technique for scoring candidate models; an algorithm for searching the candidate space and a method for interpreting the outputs of the models and their associated theories. Then in later chapters the overall approach is evaluated against empirical data, using macro level data in Chapter 6 and micro-level data in Chapter 7.

Chapter 5

Theory Evaluation, Search and Interpretation

5.1 Introduction

This chapter builds on the solutions laid out in the previous chapter, proposing approaches to the final three of the five research requirements laid out in Section 1.1.1. These requirements are entailed by the overall research objective of developing and evaluating a solution to the problem of selecting an agent based model specification and its associated theory from the many possible candidate specifications that may exist. In common with the structure of Chapter 4, this chapter references the relevant sections of Chapter 3, and draws on elements of the literature review in Chapter 2 to define the requirements of each of the solutions proposed.

The three requirements addressed in this chapter are:

3. *Theory scoring* — To address the need for the candidate models and their associated theories to be evaluated in a way that accounts for theoretical conformity and degree of fit to empirical data, Section 5.2 presents a number of ways of evaluating the qualities of the model and the associated theory. The scoring system developed can evaluate the characteristics of a theory represented in the ontological form proposed in Chapter 4.

4. *Search* — To meet the requirement to efficiently search the space of possible theories, models and their parameters resulting from the candidate theory specifications, Section 5.3 reviews the role of search in the theory selection process, identifying a genetic algorithm as a suitable means for searching the theory space and if necessary the parameter space within each theory specification.

5. *Model interpretation* — To satisfy the need to interpret a model in terms of a theory, in a way that allows an appropriate action to be selected, Section 5.4 proposes a simple recommendation process which is able to evaluate the selected theory in terms of its implications for the candidate actions that an automated agent could take.

To put these three research requirements in context, Figure 5.1 (described in Section 1.1.1) shows the area to which they apply highlighted and situated in the overall structure of the automated theory selection design.

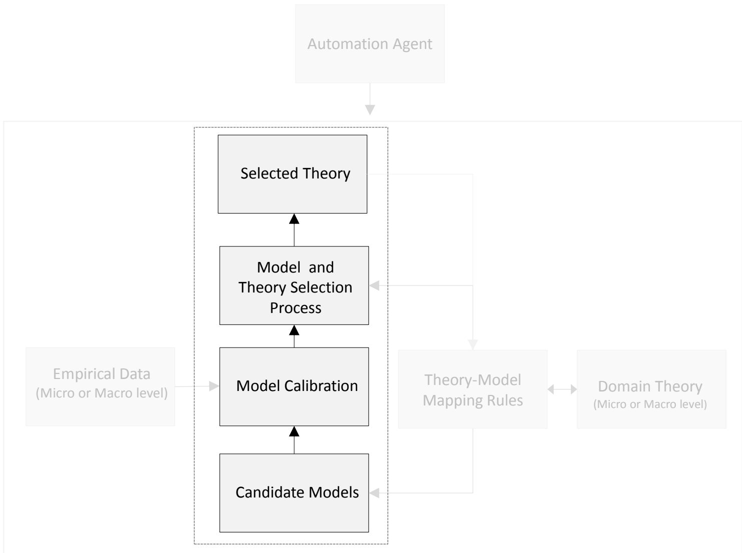

Figure 5.1: The elements of the research design addressed in this chapter

5.2 Theory and Model Scoring

In order to select a model from a set of possible alternatives, a method for evaluating the merits of the contending candidates is required. This section addresses this problem — the third of the five research requirements outlined in Section 1.2.

In Chapter 3 a form of model scoring was used by manufacturer *A* to penalise models that violated its existing beliefs about the expected direction of a relationship, and to reward models that had a higher level of fit. The details of the scoring are set out in Figure 5.3.2.1. Given that the only formal theory representation available to manufacturer *A* was the set of beliefs laid out in Section 3.3.2, there was a limited amount of reasoning that the manufacturer could carry out regarding the qualities of the theory implied by its model. This section extends the theory scoring criteria that can be applied by using the theory representation described in Section 4.2 above. As seen in Section 2.5.2, there are a large number of possible criteria that can be chosen to score a model or theory. In addition, some specific approaches to model scoring, reviewed in Section 2.6.3.1, have already been developed in an EBM context.

5.2.1 Theory and Model Scoring - Requirements

Although some of the methods used in EBM form a useful basis for developing a scoring approach, based on the review of the literature in Chapter 2 there seem to be two key reasons why a more developed model selection criteria would be needed in ABM than those already designed for EBM:

1. **It should be able to support a more complex theory representation than that used in EBM** — as discussed in Section 2.5.1 of the literature review, the task of assessing the properties of a theory

in ABM is more involved than counting the parameters or variables of which it is comprised, since theories are often deeply embedded in the design of a model's structure and involve many settings and parameters. For example, in Schelling's model of neighbourhood segregation [176] the theory of neighbourhood selection based on similarity is embedded in multiple model mechanisms, such as an agent's ability to identify its neighbour's colour and also its ability to recognise whether a diversity threshold has been reached. Altering either of these elements in isolation would leave the other as a redundant parameter. In addition, ABM implementations have sometimes been criticised for being over-parameterised, since models with realistic assumptions and agent descriptions can contain many parameters [205].

2. **It should be able to score models at multiple levels of aggregation** — as discussed in Section 2.4.4.2, multiple goodness of fit criteria may need to be considered since the degree of model fit to a micro-level data series may differ to that at a macro-level, once the micro-level data has been aggregated.

    - *Micro validation* refers to the behaviour of the individual agents — comparing each agent against empirical data or an expected outcome.
    - *Macro validation* is usually conducted against higher level aggregate patterns, often time series.

    An action that leads to an improvement in a diagnostic at micro-level will not necessarily create an improved fit at macro-level, so an automated ABM selection procedure would need to be able to dynamically calculate and trade-off different diagnostics at different levels of aggregation.

5.2.2 Theory and Model Scoring - Design

Having reviewed the requirements of a scoring system for ABM in the sections above, this section lays out a a proposed method for scoring models and theories that meets these requirements. As discussed in Chapter 4, representing a theory as an ontological system means that the theory's properties can be evaluated in a systematic fashion to assess its qualitative characteristics. Various possible criteria exist for evaluating a theory, many of which were discussed in Section 2.5.2. Some of these can be assessed as part of an automatic scoring system. In any selection process there is likely to be a trade-off between the different criteria, since any given theory may score well on some aspects but less well on others.

- **Fit** — As discussed in Section 2.5.1, a model's ability to *explain* empirical data has been central to the debate about the value of causal or explanatory models in ABM circles. Grune-Yanoff argues that if a model does not accurately reproduce the target state, there is no actual explanandum of it [85]. Model accuracy is a measure of the extent to which the model creates output which reproduces the target state. Figure 5.2 shows the point at which the degree of correspondence, or fit, is evaluated. The figure shows the policy inputs and policy outcomes generating a target time series, which the model aims to recreate by generating a time series based on the input independent variables and the

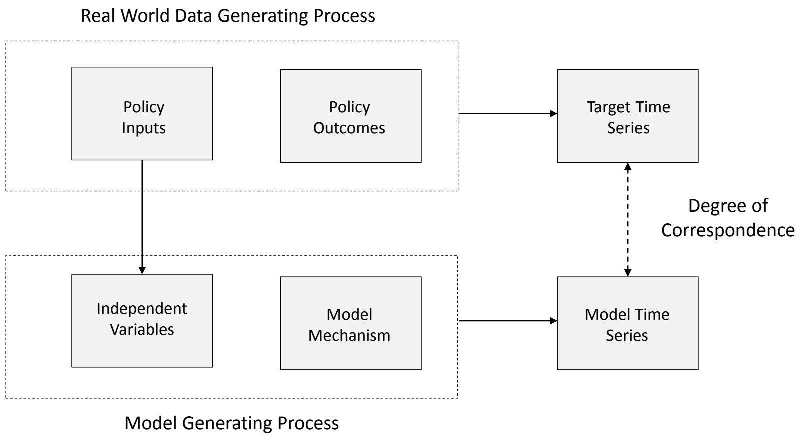

Figure 5.2: Real world and modelled data generating processes

model mechanism. The correspondence of these two time series — the model's accuracy — can be evaluated using an equation such as the model fit calculation used in Section 3.3.2:

$$ModelAccuracy = 1 - \left(\frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2} \right) \tag{5.1}$$

where y_t is the value of the target variable and $\hat{y}_t$ is the modelled value for a particular time period.

The model fit calculation above is a measure of the degree to which the predicted value of $\hat{y}_t$ is able to approximate the real level y_t . A higher level of model fit implies that the model has a greater ability to re-create the true level of the target variable. The same approach also applies to predictions of individual level events, that may not necessarily be happening over time.

In addition to model fit, the ontological structure of the theory representation described in Section 4.2.2 means that qualitative characteristics of the theory can be algorithmically evaluated and taken into account in the scoring procedure. The examples below concern the assessment of simplicity and coherence:

– **Theory Simplicity** — theory simplicity can be considered to be a measure of parsimony — a theory which contains many elements is considered less parsimonious than a theory which contains fewer elements. The value of parsimony was discussed in Section 2.4.4.2, and it was pointed out that its pursuit should not be an objective in its own right, since omitting the wrong variables may lead to models that are causally invalid. However in Section 2.5.2 it was also noted that Kuhn [119] considers parsimony to be a valid criteria for choosing between theories. In this case theory simplicity could be measured as the number of elements included in the theory, for example at a preference level:

$$TheorySimplicity = \frac{MTE}{PS} \tag{5.2}$$

where MTE is the number of modelled theory entities and PS is the total number of preference sets from which an entity could have been chosen.

In the theory design presented in Chapter 4 it was proposed that preferences can be represented as a hierarchy, with higher level preferences linked to sub-preferences. Preference sets in this context are the higher level preference groups. In other words, the more theory elements that are measured in any particular theory, the lower will be the *theory simplicity* score associated with it.

– **Theory Coherence** — coherence in this context refers to a conceptual alignment between elements in the theory, for example a coherent pair of concepts would be to describe someone as both *loving* and *kind*, but an incoherent set of concepts would be *loving* and *hateful*. Thagard argues that when integrating information about a person, we try to achieve coherence among concepts by reconciling conflicts among various pieces of data that we have about them [195]. In the context of the ontology-based theory, coherence can be considered to be the degree to which theory elements are selected from similar or diverse preference sets. This coherence can be manually coded by the user into a matrix that contains pairwise coherence scores for each element and the other possible elements:

	Theory Element					
Theory Element	1	2	3	$\cdots$	$TE-1$	TE
1	$cs11$	$cs12$	$cs13$	$\cdots$	$cs1TE-1$	$cs1TE$
2	$cs21$	$cs22$	$cs23$	$\cdots$	$cs2TE-1$	$cs2TE$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
TE	$csTE1$	$csTE2$	$csTE3$	$\cdots$	$csTETE-1$	$csTETE$

The table above shows the elements of the theory representation, 1 to TE (the number of theory elements) on the horizontal and vertical axes. Each cell, cs contains the degree of conformity between the elements identified on each axis, scaled between 0 and 1. So for example the degree of conformity between theory element 1 and theory element 2 is contained in cell $cs12$, and if the cell contained the value 1 that would indicate a high degree of coherence between the two elements.

To arrive at an overall coherence score, $TheoryCoherence$, that takes into account all of the represented theory elements (excluding those that are not part of the learnt specification), the mean of all active coherence scores can be calculated to give an overall level of coherence in the theory.

- **Overall Model Score** — As mentioned in Section 1.1, choosing a model can be a complex task because of the trade-offs between theory and empirical evidence, and contending success criteria. An overall score for a particular theory can be calculated that incorporated both its quantitative fit and its qualitative properties, for example the overall score for each theory can be calculated as the weighted sum of the individual scores, for example as:

  $$OverallTheoryScore = AW * ModelAccuracy + SW * TheorySimplicity + CW * TheoryCoherence$$

  The $OverallTheoryScore$ is therefore a weighted combination of the set of scoring criteria applied. As an example, if the $ModelAccuracy$ criteria was valued above the $TheoryCoherence$ score, the weights given to each of the factors AW and CW can be adjusted to reflect their perceived importance.

Once calculated, the overall score for each theory and model combination can be passed back to the search process, described in Section 5.3, as the basis of the model ranking procedure.

5.3 Search Mechanisms

In order to efficiently select a model from the set of possible models, a method is needed that can explore the range of candidates permitted by the theory specification. This section addresses this, the fourth of the five research requirements outlined in Section 1.2 — the requirement to efficiently search the space of possible theories, models and their parameters resulting from the candidate theory specifications. Then in Section 5.3.2 various methods are considered that could be used to meet the requirements and an appropriate solution for this task is identified.

This section builds on the work on search procedures discussed in Chapter 3, in which two search objectives were pursued — one (detailed in Figure 5.3.2.1) a search of variable combinations that could be validly combined in an equation, and the other (detailed in Figure 3.3 a search of parameter values that could fulfil the requirements of fit maximisation while maintaining theory compliance. The search process in this section is considered in the broader context of automated theory selection rather than the parameter and variable selection approaches applied in Chapter 3. The primary difference is that there may be an extended range of elements that need to be selected such as utility functions and social network topologies.

5.3.1 Search Mechanisms - Requirements

Because of the complexity of many agent based models, there are a number of factors that need to be considered in the search process. Some of the specific issues revealed by a review of the literature include:

1. **It should be able to accommodate a variety of search objectives** — There is a wide variety of entities that may form part of the search space. For example, part of the theory search process may be to determine whether average daily temperature creates a better fit to the observed data than maximum daily temperature. At the environment level the process might need to be able to select such elements as the interaction networks which connect the modelled agents, and their topology. At the model level the selection process would need to be able to vary the initial state of the environment and agents, as well as have the ability to control the visibility of global and local variables. Based on the overall structure of the theory selection approach, the automation agent needs to perform two types of search:

   - **Theory search** — a high level search of the candidate theory space, defined by the theory specification described in Section 4.2.2.1. This may contain elements such as the social network topology and the form of the utility function used by the agents.
   - **Parameter search** — The selection procedure also needs to calibrate the associated model under evaluation by conducting a search for the best value for each parameter. The type of search involved depends on the form of the model with which the theory is associated — and in some cases parameters might be wholly determined analytically or by extraneous information. There is considerable existing work on parameter estimation in equation based models, some of which was reviewed in Section 2.4.1. Work on calibration in ABM is at an earlier stage than that in EBM — much of the research carried out so far using ABM has been aimed at conceptual exploration, while relatively little work has been published with the aim calibrating models against empirical data [92, 91]. Some of the reasons for the lack of empirically oriented research may include the more qualitative focus of the researchers currently using ABM, difficulties in obtaining suitable data, and a lack of established methodologies for validating models designed for predictive accuracy. However, in common with EBM, ABM calibration generally follows one of three broad routes: estimation from data; imposition of external information; or a combination of the two approaches [71, 209].

2. **It should be efficient** — for both parameter and theory search, the strategy that is most likely to reveal the best theory involves an exhaustive search of the solution space, examining all possible

candidates [27]. However this quickly becomes impractical when there is a large number of possible components. For example, in a model with N candidate inputs there are $N!$ possible paths, each of which may have up to N steps [94], and with a set of ten parameters, with each one able to take ten possible values, there may be ten billion permutations [194]. Assuming that not all possible models can be reviewed, some kind of selective search is required. In common with EBM, the computational requirement can quickly become very large since within any given candidate theory there is a range of possible parameter values.

3. **It should be able to avoid path dependence** — Any model selection strategy which involves sequential actions introduces a form of path-dependence, and automatic model selection processes have been criticised for their susceptibility to this problem [94]. In an ABM this path dependence would also influence, for example, the selection of a particular decision rule, given that another rule or network topology is already active.

4. **It should be able to avoid converging on local optima** — One of the issues that arises when dealing with complex models is that hill-climbing search strategies can be sub-optimal, potentially reaching a local optimum and stopping [141], while approaches that vary one model element in isolation of the others may ignore their underlying connections and mis-assign causality. Because of the potentially complex pattern of cross-correlations between factors, and the other trade-offs between model components, the selection process has no complete map of the selection space. When using correlated data, the components already included in a model can affect its current diagnostic criteria, as well as the sign and significance of the remaining candidate components. In an ABM this path dependence would also influence, for example, the selection of a particular decision rule, given that another rule or network topology is already active.

5. **It should be able to handle emergent and adaptive behaviour** — modelled agents may respond to a certain situation by adopting a different rule rather than by varying a parameter, creating a highly non-linear response. Moss and Edmonds give the example of changes to the rules governing financial markets after the 1929 crash [150]. In addition the process may need the ability to test for changes in agent parameters over time.

5.3.2 Search Mechanisms - Design

In Section 2.4.2 various issues and approaches were reviewed relating to verification in agent based models, including methods for searching parameter spaces. Edmonds and Bryson argue that the complex nature of emergent patterns and non-linear interactions between parameters mean that the outputs of ABMs can rarely be described by mathematical functions, and that analytic methods usually prove insufficient [52]. In this section existing research is drawn on in considering which of the variety of search algorithms that have been developed could be employed to search the hypothesis space, including grid search methods, gradient descent, simulated annealing and genetic algorithms [172].

Kennedy [109] argues that most optimisation procedures aim to minimise or maximise an objective function. Grid search methods involve calculating the value of the objective function across a set range of parameters, and are considered to be inefficient in that they involve evaluating every combination of input parameters [109]. Faster methods are usually sought, which typically involve selecting a set of starting values for the parameters, and iteratively calculating the objective function then moving the parameters towards a direction in which the objective function is increased. The iterative search is concluded when a pre-defined convergence criterion is met. Hill climbing optimisation methods are an example of this, consisting of a loop that continuously moves in the direction of increasing value — uphill [172]. In their study of calibration using a variety of search algorithms, Stonedahl and Wilensky [188] found that genetic algorithms are an effective means of exploring the parameter space of ABMs. But they note that it is unclear what circumstances favour the use of a genetic algorithm over hill climbing search mechanisms. Fabretti found that in her research the Nelder Mead simplex algorithm sometimes failed to find the global optimum and the procedure had to be restarted, while this did not occur when using a genetic algorithm which she found to be more robust when applied in a noisy environment [61].

In addition various researchers have found that hill climbing methods may be subject to failure because the kind of objective functions found in ABM do not always behave well enough to guarantee a global optimum solution [61]. Russell and Norvig concur that hill climbing does not look beyond the immediate neighbours of the current state and often fails to find a goal when one exists, because they can get stuck on local minima [172]. Similarly, in his comparative study of genetic algorithms and hill climbing approaches, Michaelwicz argues that hill-climbing produces local optima only, and these optima depend on the starting point of the search [142]. Genetic algorithms have been used successfully in a number of applications [60, 198, 194, 32], and offer a relatively efficient solution to the theory search required in this application. Stonedahl and Wilensky conclude that in general, the prospects seem bright for using genetic algorithms to improve model exploration and analysis.

One of the central features of a genetic algorithm is the representation of each population member. Traditionally, binary strings have been favoured by many GA researchers, but there have also been successful implementations using non-binary representations [14]. In the context of theory search, the standard representation of 0s and 1s can produce illegal solutions. For example if the three possible choices of social network were each represented by a 0 or 1 in a chromosome, a mutation or crossover operation could create a chromosome in which two social network specifications were active. Since the choice of social network in the model is a *one of* rather than an *any of* representation, this chromosome would represent an illegal solution. One solution to this problem is to introduce a repair algorithm to modify illegal chromosomes so that they can be evaluated. However, Fogel [63] concludes that there is no particular advantage in using binary representations, and Michalewicz suggests that in reality most researchers modify their implementations of genetic algorithms by designing specific genetic operators to suit the problem to be solved [142]. For this

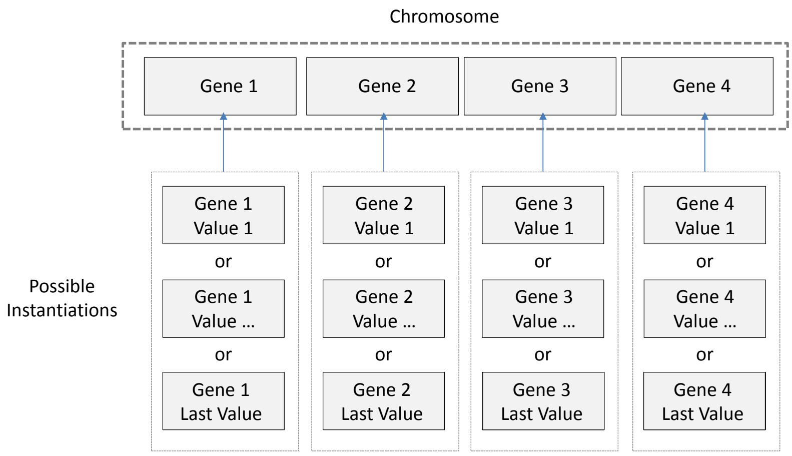

Figure 5.3: The non-binary chromosome representation used in the genetic algorithm

application a non-binary alphabet seems appropriate, due to the exclusivity of some of the elements in the theory sets. The representation is shown in Figure 5.3. Each gene in the chromosome can be instantiated with any of the possible values that are relevant to that particular group. A gene represents either the value of a numeric parameter or a categorical value, and the element that replaces it as part of the crossover or mutation operations are part of the same group, for example an alternative numeric parameter, or another social network specification. Figure 5.4 shows how the cross-over process works with the non-binary alphabet, and Figure 5.5 illustrates the role of the mutation process. In each case, the value contained in each gene can only be replaced with one of the other values that are relevant to that gene. Figure 5.6 shows the workings of the genetic algorithm over a number of generations.

The details of the algorithm's parameters are purposefully left unspecified since there is a large variety of possible contexts that the procedure needs to be able to cover. In principle any element of the theory representation could be considered a part of the searchable space, including the utility function used by the agents, the topology of the social network, and any number of constants that might control elements such as the period over which the simulation is to run.

The procedure in this form can search both continuous and categorical spaces, and can be modified to do both simultaneously, for example if the parameter and theory spaces were to be explored simultaneously rather than hierarchically as they are in this example. In other words, rather than carrying out the operation *Calibrate parameters within each model* for each model as a sub-task, the parameter space could be brought up into the overall theory space being searched by the higher level process.

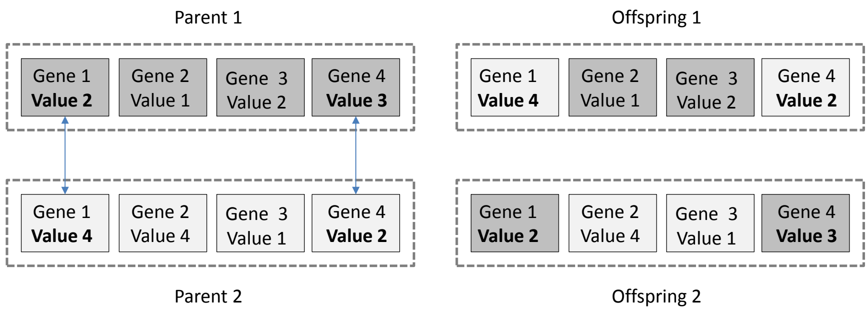

Figure 5.4: The non-binary crossover process used in the genetic algorithm

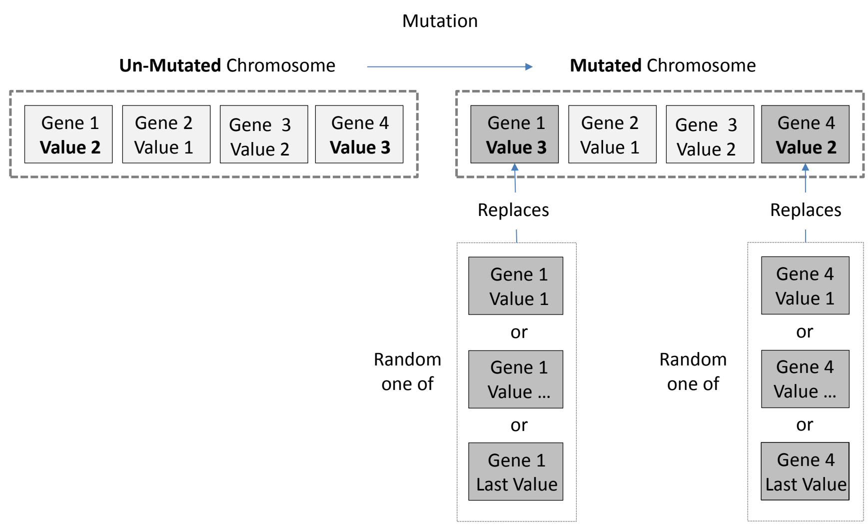

Figure 5.5: The non-binary mutation process used in the genetic algorithm

Figure 5.6: General theory and model framework search algorithm

1. *Create Initial Population of theories with random elements drawn from permissible ranges defined by relationship rules (outlined in Section 4.2.2.1)*
2. *Repeat until termination criteria met*
   (a) *For each model*
      i. *Calibrate parameters within each model*

   (b) *Score models and associated theories against set criteria (discussed in Section 5.2.2)*
   (c) *Create new population of models. Reproduce models proportionally to score*
   (d) *Random crossover between model pairs*
   (e) *Random mutation on new model population*
   (f) *Check for termination criteria — whether model and theory score acquired greater than target or iterations exceeded*
   (g) *Loop, if not terminating*
3. *Exit procedure if terminating*

Table 5.1: Performance of the algorithm over 30 trials

Parameter	Occasions correct	Occasions incorrect
Advertising	28	2
Promotions	28	2
Coupons	28	2

5.3.2.1 Exploration of a genetic algorithm's performance

Stonedahl and Wilensky [188] point out that consideration should be given to the treatment of model stochasticity and noisy objective functions. To test the stability and accuracy of genetic algorithms in an ABM context, the search procedures used by Manufacturer *B* in the agent simulations from Chapter 3 were tested under different levels of search iteration. The details of the genetic algorithm used are detailed in Figure . The models were re-run 30 times, using a level of 50% social interaction — in other words about half of the maximum level tested in Chapter 3. To illustrate the progress towards stable estimates in a single run, Figure 5.7 shows the improvement of the score criteria for a single iteration of the GA as it proceeds through 100 generations. Figure 5.8 shows the corresponding parameter estimates for the same iteration. In this particular iteration convergence was reached after 18 generations, and since the best possible model had already been achieved there was no further improvement over the subsequent 82 generations. To test the stability of the estimates over a larger number of iterations, the process was run 30 times. Table 5.1 shows the number of cases in which each of the parameter estimates was exactly correct. Because the algorithm is running on synthetic data, the actual parameters used by the agents are known, and can be compared with the estimated parameters.

In 93% of the 30 iterations, the genetic algorithm achieved a solution which was exactly correct. In the 7% of the cases when the exact solution was not obtained, the estimates were very close, as Figure 5.9 shows. The true parameter for the advertising weight is 20, and 10 for both coupons and promotions. The results from this evaluation suggest that the genetic algorithm approach is consistent and accurate in achieving the correct solution, where the correct solution is known.

5.4 Model Interpretation

Once a model and its associated theory has been selected from the range of possible candidates using the methods proposed in the sections above, an automated approach needs a way of interpreting the implications of the selected candidate. This section addresses the fifth of the five research requirements outlined in Section 1.2 — the need to interpret a model in terms of a theory, in a way that allows an appropriate action to be selected.

In the simulation experiment presented in Chapter 3 the two agents *A* and *B* used a form of interpretation to select the actions that they would take in each period, as defined in Section 3.3.4. In the simulation, agents *A* and *B* selected the action associated with the coefficients from their models which implied the

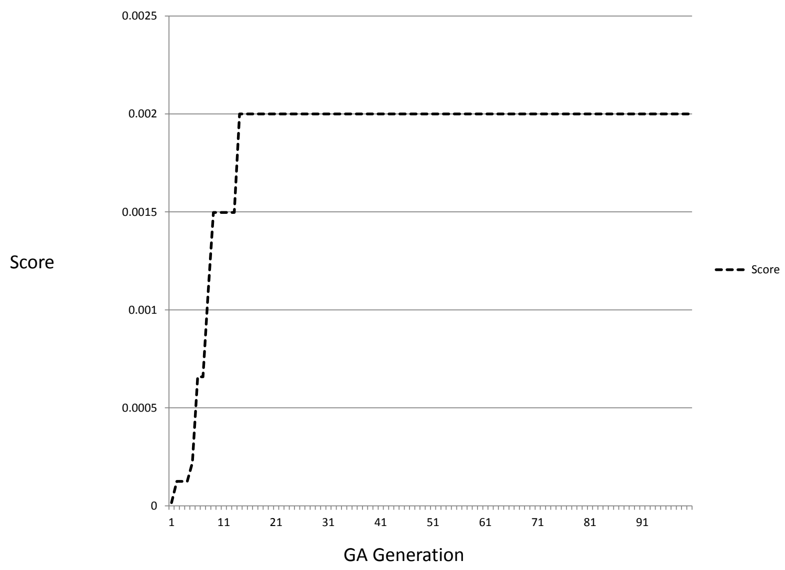

Figure 5.7: The development of the calculated score over 100 generations of the genetic algorithm

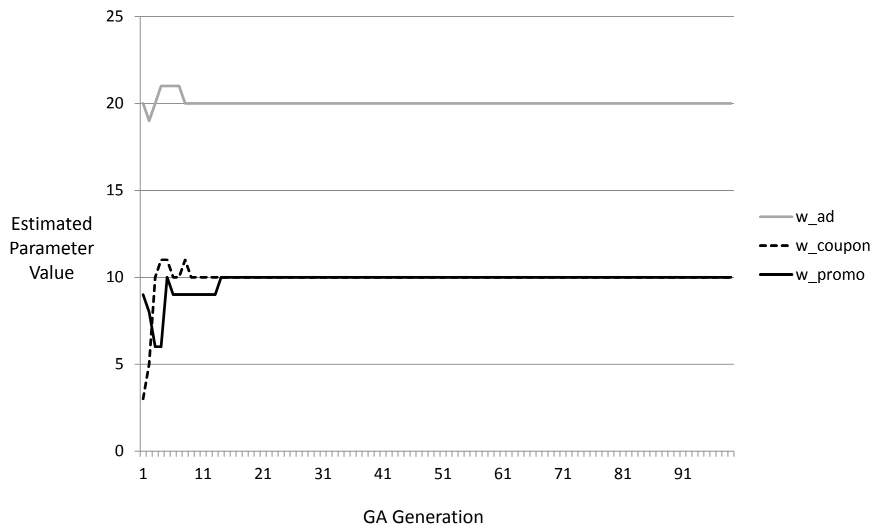

Figure 5.8: The development of the estimated parameters over 100 generations of the genetic algorithm

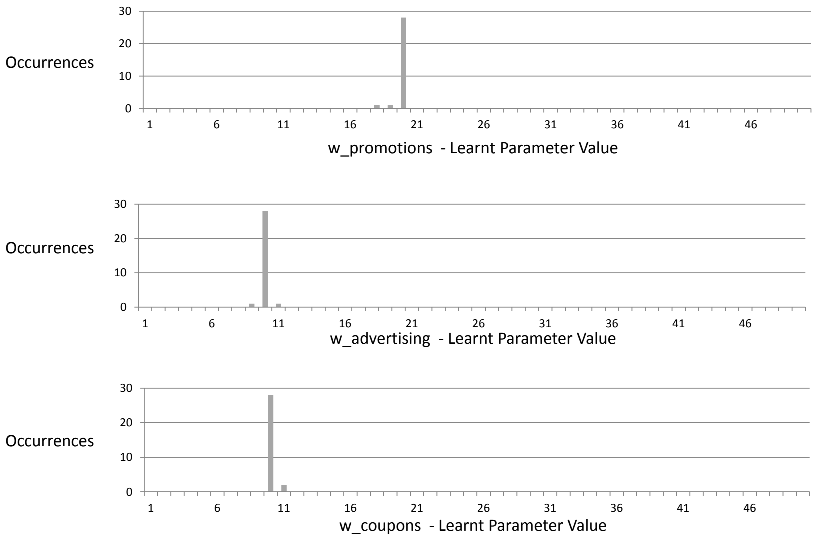

Figure 5.9: The distribution of parameter estimates over 30 iterations of the genetic algorithm

highest level of return. In this section a simple approach is outlined for determining which actions an automated agent might associate with a discovered theory element, building on the method already deployed in Chapter 3 by explicitly incorporating the possibility of uncertainty about the quality of the input information, as well as the possibility that there is uncertainty about whether the rule itself is valid, or necessarily implies the best course of action.

5.4.1 Model Interpretation — Requirements

The following requirements were suggested by the literature review, documented in Chapter 2:

1. **It should be able to incorporate information about uncertainty in the input data** — In Section 2.5.1 various issues were considered about the degree of certainty that could be ascribed to theories learnt through an agent based modelling approach. For example Grune-Yanoff [85] argued that in the case of Epstein's Anasazi model, the behavioural rules could not be verified because different rules create similar fits. Similarly, in Section 2.5.4, Goldfarb *et al.* argued that an indefinite number of rival theories may be created that have identical empirical consequences.

2. **It should be able to incorporate information about uncertainty about the quality of the implied recommendation** — Practitioners in different domains may have varying levels of confidence that a specific policy choice is the correct response to a particular empirical fact, even if they have a high degree of confidence that the fact itself is the case.

5.4.2 Model Interpretation — Design

The factors that would need to be taken into account in assessing accuracy may vary substantially across applications, making it difficult to define an exact process. Probability elements could be used to apply weights to the rule recommendations. Rules could be filtered to eliminate those that are not sufficiently certain, and sorted based on the expected utility of the outcome to select the most attractive action to take.

The model interpretation method proposed takes the outputs from the model and theory components and infers an optimal action to take. The proposed system uses a set of IF THEN rules to match elements from the induced theory to the optimal recommendation. These rules are specified by the designer to associate different theory outcomes with relevant courses of action. For example if a particular preference is discovered to be A as part of the selection process, then an associated rule might be *'If preference =A then action = Advertise A'*.

Where sufficient appropriate information was available, certainty scoring could be based on the MYCIN methodology [99], which takes into account the combined probabilities of the antecedents in the rules. The methodology proposes that the certainty about the facts upon which the rules are based should be collectively assessed and the minimum certainty about any of the data in the rule should apply to the whole rule. So for example, a theory might be learnt that suggests that $x = A$ and that $z = B$. There may be uncertainty about how accurate the theory elements are, and given the rule *'if x is A and z is B then y is C'* and a probability of 50% that x is actually A, and 30% that z is actually B, then the certainty of the rule would be 30%. In other words the overall antecedent score is the minimum of all the associated certainty scores. $CFAntecedents = Minimumof(AllCertaintyFactorsRelevantToRule)$ where $CFAntecedents$ is the certainty that the antecedents in the rule are correct.

From a statistical perspective, assessing the probability that the learnt theory element is correct would require some kind of maximum likelihood or variance decomposition method. There is some existing work in this area, including the theory of maximum likelihood approaches in agent based simulation [115] and its applications in computational finance [136, 4, 127]. This areas is discussed further in Section 8.4.

As well as levels of certainty about the inputs to the rule, there may be degrees of confidence in the outcome that a rule recommends. For example the designer of the rule *'if x is A and z is B then y is C'* may only be 50% confident that under those conditions y is C. Combining uncertainty about both the antecedents and consequent, the following formula can be applied to assign each rule an overall certainty score:

$$CFoutcome = CFAntecedents * CFrule. \tag{5.3}$$

where $CFoutcome$ is the certainty associated with the interpretation overall, $CFAntecedents$ is the certainty associated with the theory element, and $CFrule$ is the certainty of the rule designer that the rule is correct.

The overall certainty of the rule is therefore a combination of certainty about the inputs and outputs.

To return to the example used above, since the minimum antecedent certainty was 30%, and the probability that the rule would hold is 50%, then the overall score for that rule woule be 15%. Rules can be filtered to eliminate those that are not sufficiently certain, and sorted based on the expected utility of the outcome to select the most attractive action to take.

5.5 Conclusions

This chapter addresses the final three of the five research requirements outlined in Section 1.2, the previous two having been addressed in Chapter 4. In this chapter:

3. In Section 5.2.1the characteristics that would be needed from an evaluation methodology were considered, then in Section 5.2.2 a method was proposed for evaluating the quantitative fit of the model and the properties of the associated theory — using evaluations run against the ontological theory structure.

4. In Section 5.3.1 some of the particular features that the search mechanism needs to be able to handle were reviewed, and in Section 5.3.2 a genetic algorithm was identified as a suitable means for searching the theory and parameter space. Finally, the need to interpret a model in terms of a theory was considered.

5. In Section 5.4 a recommendation process was proposed which can evaluate the selected theory in terms of implied recommendations for an automated decision system. A method is also proposed for weighting the implied recommendations based on the level of certainty associated with them.

Combining these three elements with the two areas discussed in Chapter 4 creates the foundations of an agent based model selection system intended to solve the primary research problem of selecting an agent based model specification from a set of possible candidates. In combination the approach involves an agent-centric theory representation, a method for mapping a theory to a model, a proposed technique for scoring candidate models, an algorithm for searching the candidate space, and a rule-based solution to interpreting the outputs of the models. In the following chapters, Chapter 6 and Chapter 7 the approach is evaluated against real data.

Chapter 6

Theory Selection with Macro-Level Empirical Data

6.1 Introduction

This chapter presents an applied evaluation of the solution proposed in Chapters 4 and 5 to the research problem defined in Chapter 1 — selecting an agent based model specification from a set of contending candidates. The previous two chapters were concerned with establishing the conceptual and methodological framework necessary to solve the problem and in this chapter the proposed solution is evaluated using an empirical data set. The information in this dataset has been retrieved and stored at a macro level — in other words each record represents an aggregation of individual behaviours rather than the behaviour of a specific individual.

The aim of the chapter is to evaluate the success of the proposed approach when applied to a real dataset at a macro level, and to compare its performance against existing methodologies. The chapter is divided into five broad sections. In Section 6.2 an overview of the domain — online social networking — is given, with a description of the social sharing processes that were active in the real, *target*, social network process. In Section 6.3 the theoretical structure used in this theory selection process is defined, along with the elements contained in it. In Section 6.5 the elements of the theory defined in Section 6.3 are related to the agent based model to be used in this analysis. The model structure and simulation process are defined in terms of the agents characteristics and behaviours. In Section 6.6 the data, sourced from an online social networking site, are described. Finally the results and management recommendations from the study are presented and discussed in Section 6.7.

6.2 Domain Background

This section gives a high level overview of online social networks and the way that they are used by companies and consumers. As mentioned in Section 2.6.2, marketing is one of the methods that companies use to reach their goals and objectives. As part of their marketing activities, many companies use posts on their social media brand pages with the aim of building relationships with their customers, and a large proportion of social media users follow brands on social media [46]. Companies are able to post text, images and video to these pages which are then visible to their followers. Followers can interact with these posts by *liking* or commenting on them, and the social media site may then transmit the fact that a follower has interacted with a post to other users in the follower's network, creating a form of social advertising. The diffusion effect for brand posts on the social network is enabled by the *news feed*, which appears on a user's homepage and shows them recent activity related to their friends, including amongst others *likes* for entities on the site, profile changes, comments, and interaction with applications. The *news feed* allows for active and passive sharing — active sharing is made through posts, while passive sharing involves reporting of activities. The collection of available news is filtered by the social network and a subset is presented to a particular user.

A user may feel evidence of increased social pressure where multiple friends in their network are performing the same actions, for example *liking* the same brand post or the same comment on a brand post [190]. Various studies have looked at the role of these cascade effects in online social networks [210, 35]. Watts and Dodds distinguish between local and global cascades, with global cascades occurring only when a *critical mass* is reached amongst the early adopters — agents who are activated after being influenced by just one neighbour. They find that the development of cascades is more dependent on the overall network structure than on the influence of any agent, but also that influential users, those with a larger network, are more likely to trigger larger cascades.

Figure 6.1 shows the process that was in operation in the social network during the period that the data set covers. The figure shows posts originating from updates to the *brand page* and being transmitted to the direct audience — essentially users who subscribe to the page. Users who *like* the post create a secondary transmission to users in their social network — who may or may not themselves be subscribers to the *brand page*. Users who are already subscribers will see the post an additional time, while users who are not become the *viral* audience — users who see the post because of sharing rather than direct transmission from the brand.

In this section an overview was given of the domain that the theory and model selection is to be applied to. In the next section a theory representation is presented that encompasses the main elements of the domain.

6.3 Theory Specification

In this section the theory representation to be used in this application is defined, incorporating as far as possible the elements highlighted in the domain overview above. Section 4.2.2 outlined a generic design for a theory representation that would be suitable for automated agent based modelling processes, and in this section the elements of that design are related to the specific domain and data set under study here. Figure 6.2 shows an overview of the theory specification that is used to reflect the theoretical environment, which is made up of exogenous factors, agent behaviours and characteristics, and emergent factors. The

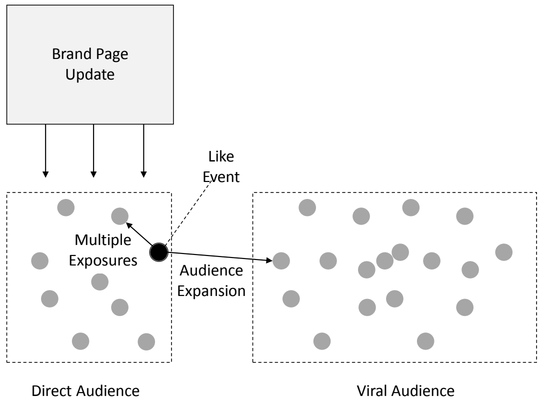

Figure 6.1: Propagation of posts from direct audience to viral audience

factors shown in Figure 6.2 are described in more detail below.

6.3.1 Exogenous Factors

Exogenous factors are elements that come from outside of the model and can have an influence on the model's internal dynamics. As argued in Section 4.2.1, since any model is a partial representation of the real world, the theory framework should be able to incorporate external inputs. Table 6.1 shows the exogenous elements that are reflected in this particular theory. These consist of various properties of the brand post — text and/ or image based features which are made available by the advertising company to users on the social network. In addition, two climatic factors are considered which may have an impact on web usage, as well as PR activity by the advertiser.

6.3.2 Relationship Rules

To recap the definition laid out in Section 4.2.1, relationship rules in this context define the hypothesis space of relationships to be tested, laying out the set of hypotheses about how agent preferences map to exogenous and emergent factors. In this case existing academic studies provide few indications about the range of impact that exogenous factors may have, but give some guidance on whether the impact is likely to be positive or negative. Some relationship rules are therefore set to allow any parameter value between -1 and 1, while others, like social pressure, are set to take any positive value. Other parameters are allowed to take any negative value. Because there is no defined usage rate the purchase cycle is set to 1, meaning that any agent has the possibility of *liking* a post once for any *brand post* event.

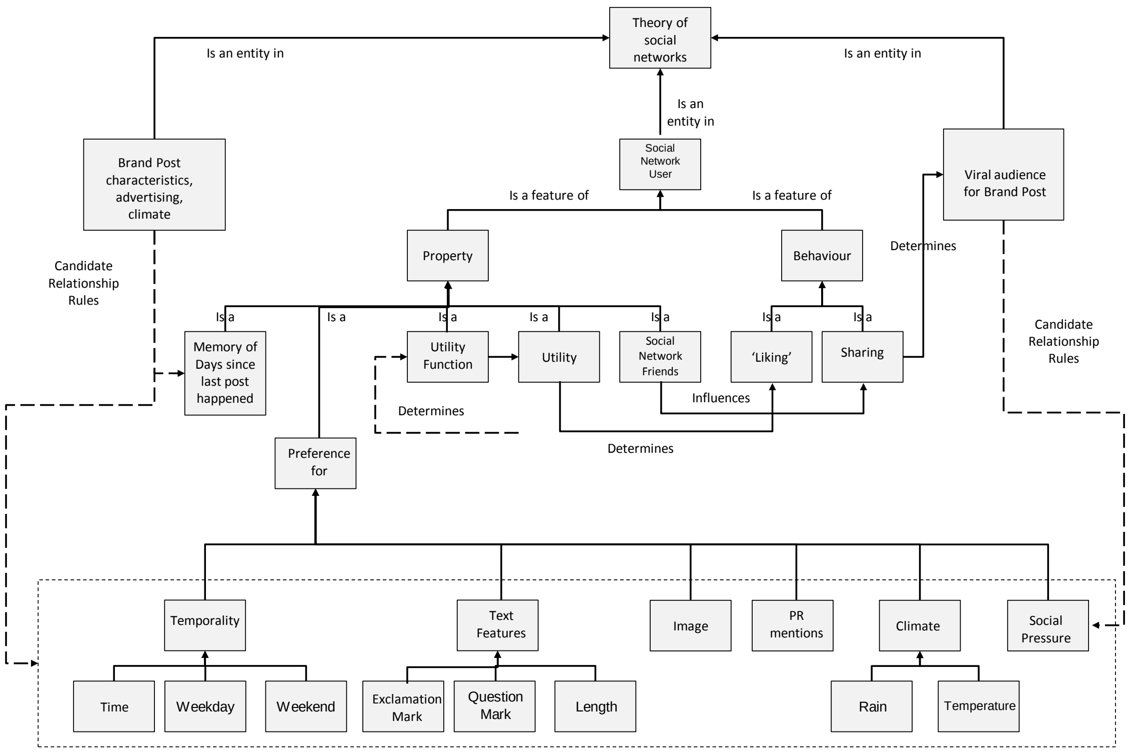

Figure 6.2: The online social network theory specification

Table 6.1: Exogenous elements considered in the model

Variable Name	Variable Label	Variable Description
exImagePresent	Image present	A binary variable which takes the value of one if the post was accompanied by a photo or other image and zero otherwise.
exTextPresent	Text present	A binary variable which takes the value of one if the post was accompanied by text, and zero otherwise.
exQuestionMarkPresent	Text contains question mark	A binary variable which takes the value of one if the post contained a question mark ('?'), and zero otherwise.
exExclamationMarkPresent	Text contains exclamation mark	A binary variable which takes the value of one if the post contained an exclamation mark ('!'), and zero otherwise.
exLengthOfPost	Length of post	A continuous variable which reflects the total length of the post in terms of the number of characters used.
exProductMentioned	Product mentioned	This is broken down into three sub-variables, one for each of the possible product groups for sale that could be mentioned. Each of these variables takes the value of one if the respective product is mentioned, and zero otherwise.
exPostedAtWeekend	Posted at weekend	A binary variable that takes the value of one if the post was made on a Saturday or Sunday and zero otherwise.
exPostedOnWeekday1, exPostedOnWeekday2, exPostedOnWeekday3, exPostedOnWeekday4, exPostedOnWeekday5	Posted on one of weekdays	This is broken down into five sub-variables, one for each day of the week. Each of these variables takes the value of one or zero.
exTimePosted	Time posted	A continuous variable, expressed in terms of the 24 hour clock, such that lower values are earlier in the day.
exTemp	UK temperature	A continuous variable, measured in degrees celsius, representing the average rainfall on the day of the brand post.
exRainfall	UK Rainfall	A continuous variable, measured in centimetres, representing the average temperature on the day of the brand post.
exPRCoverage	PR coverage	A count of the number of number of news media articles in which the brand was mentioned in the week of the brand post.

Table 6.2: Other variables considered in the model

Variable	Description
Social Pressure	The number of times the agent has seen the particular *brand post* shared
Memory- Days since last post	This is a continuous variable which reflects the number of days prior to the respective post that the last post was made.

6.4 Mapping Rules

Figure 6.3 shows the mapping rules used to connect the theoretical specification to the agent based model. The first group of items, starting with ex, represent the exogenous variables that are active candidates in the simulation. The second group, starting with $pref$ represent the agent preferences that these correspond to — these correspond with the w_{ijk} parameters defined in Section 6.5.5. The third group of entities, for example $0 : 1$, represents the sign restrictions on the utility weights that can be estimated for these parameters. In the rest of the rule the utility function is specified as being additive, the three social network candidates explained in Section 6.5.1 are laid out in an *OR* group, the social pressure function is defined as *neighbour-pressure* (defined in Section 6.5.1) and the time period is specified as being the whole period that the exogenous data covers.

Figure 6.3: The mapping rules used to connect the agent based model with the theory parameters

([exImagePresent, exTextPresent, exQuestionMarkPresent, exExclamationMarkPresent, exLengthOfPost, exProductMentioned , exPostedAtWeekend, exPostedOnWeekday1, exPostedOnWeekday2, exPostedOnWeekday3, exPostedOnWeekday4, exPostedOnWeekday5, exTimePosted, exUKtemp, exUKRainfall, exPRCoverage],[prefImagePresent, prefTextPresent, prefQuestionMarkPresent, prefExclamationMarkPresent, prefLengthofPost, prefProductMentioned , prefPostedAtWeekend, prefPostedOnWeekday1, prefPostedOnWeekday2, prefPostedOnWeekday3, prefPostedOnWeekday4, prefPostedOnWeekday5, prefTimePosted, prefUKtemp, prefUKRainfall, prefPRCoverage],$[0 : 1, 0 : 1, -1 : 1,, -1 : 1,, -1 : 1,, -1 : 1,, -1 : 1,, -1 : 1,, -1 : 1,, -1 : 1,, -1 : 1, -1 : 0, -1 : 0, 0 : 1, 0 : 1]$,*[utility-additive],[social-network-1 or social-network-2 or social-network-3],[neighbour-pressure],[full-period],[ ])*

Having defined the theory representation and mapping rules to be used, the next section lays out the agent based model to be used in this application.

6.5 Model

In the previous sections the theory structure to be used in this application was defined along with the relevant mapping rules, and this section details the elements of the agent based model that are mapped to that theory structure. The following points provide an overview of the model and its key features:

- A population of 10,000 agents was created, with each agent representing a social media user — in this case a user of the online social network service.
- Each agent has an immediate social network — a group of other agents to whom each agent is directly linked. The representation of this social network is described in Section 6.5.1.
- The agent has one action — the ability to *like* a post — which corresponds with the real world act of registering the fact that an individual *likes* the brand post by clicking a *like* button associated with the post. A *like* event by a particular agent causes the *liked* entity to become visible to the other agents in the *liking* agent's social network, with a given probability. This reflects the real world propagation dynamic outlined in Section 6.2. More detail on the agent's actions are given in Section 6.5.2. The agent's decision to like a post is based on whether or not it is exposed to the post and its utility — defined below.
- The agent has a certain level of utility for the post, calculated from the weighted preferences for its characteristics. The agents' utility calculation is laid out in Section 6.5.5.
- Each agent has a memory of previous brand posts that he has been exposed to. The memory is described in Section 6.5.4.
- The agent is able to perceive the *brand post* with its respective characteristics, and also the count of previous *like* events — actions by other agents — which are shown alongside the post itself.
- The agents in the model operate within a relatively simple environment with no spatial element.
- Initial conditions of the model are preset and it executes over a pre-specified number of iterations – the length of the historic data period. Each iteration represents a specific *brand post* event.

Having outlined the key characteristics of the model, the next sections provide more details on each of its features.

6.5.1 Social Pressure

This section describes the workings of the social pressure process used in the model. The agents in this simulation use the same social pressure mechanism as that described in Section 3.5 — each agent has a threshold for social pressure, based on the percentage of adopters in his immediate network and decides whether or not to communicate with other agents, based on the likelihood to communicate passed to it from the theory specification. As in Section 3.5 their experience of social pressure is calculated as:

$$SP_{ij} = \frac{N_j}{N} \tag{6.1}$$

where SP_{ij} is the social pressure felt by agent *i* to *like* the post *j*, N_j is the number of agents with whom agent *i* is linked who have already *liked* the post *j*, and *N* is the total number of agents to whom agent *i* is linked.

In other words, as the proportion of the agents who are linked to agent i and have already *liked* the post j increases, the social pressure, SP_{ij}, on agent *i* to *like* the post also increases.

The agents were arranged in a network based on Watts and Strogatz [202] small world. As discussed in Section 4.3.2 one of the possible theory elements that could be selected from is the structure of the network. A small world network contains a mixture of short paths connecting most of the individuals within each clique, and longer paths connecting the cliques. Three different variations of the network were tested in the calibration process, reflecting:

1. *A higher concentration of longer paths between cliques — on average 0.16 long paths per agent*
2. *A moderate concentration of longer paths between cliques — on average 0.11 long paths per agent*
3. *A low concentration of longer paths between cliques — on average 0.08 long paths per agent*

The objective of including multiple possible network configurations was to determine which network structure was most likely to be active in the target process. The hypothesis was that changing the topology of the network would have an effect on the shape of the modelled variable, and therefore the fit against the target variable.

6.5.2 Actions

Earlier, in 4.2.2.1, actions were defined as the capabilities an agent can execute to interact with the world [191]. In this simulation the agents have only one potential action — the ability to *like* posts. They decide whether to take this action based on whether their utility (described in Section 6.5.5) exceeds a threshold.

6.5.3 Perception

Agents are able to perceive the properties of the *brand post* and additional information contained in the exogenous data. As a pre-condition to evaluating a post and taking this action, agents need to be exposed to the post, however not all of the agents see every post when it is placed by the brand. Some agents may miss it when it is present on their *news feed*, either because of the frequency with which they check their account, or because of other contending information in the feed. Each agent has a probability of seeing any particular brand post naturally — without any propagation by contacts in their social network. It is assumed that agents who see the post make up the natural audience for each post event. In addition each agent has additional opportunities to be exposed to the post if there is a *like* by another agent in their network. Each agent checks their *news feed* once for each *brand post* event.

6.5.4 Memory

The role of memory in consumer behaviour was considered in Section 2.6.1, particularly as an influence on repeat purchases. It was also seen in Section 2.2 to be an important element in several agent cognitive architectures. In this simulation, memory plays a potential role in shaping agent's propensity to act in each period. Agents in the model maintain a memory of posts seen in previous periods, and in each period they decays their memory of previous posts at a given rate, to be learnt as part of the model calibration process. The equation below shows the mechanism through which each agent processes its memory of previous posts, the agent is constantly forgetting previous posts, with the agent's memory of those previous posts decaying at the rate of λ:

$$A_t = T_t + \lambda A_{t-1} \tag{6.2}$$

where A_t is a *brand post* exposure experienced by an agent at time t, T_t is the value of the *brand post* variable at time t and λ is the decay parameter.

6.5.5 Utility

This section presents the utility function used by the agents in the simulation. In Section 4.2.2 the need was discussed for a utility representation that would allow an agent to weigh up the relative merits of different sets of attributes. In this application, each agent is given a threshold in the simulation for *liking* any particular brand post, based on the combination of characteristics, weighted by preference, that the post contains. This threshold exists irrespective of whether or not the agent actually comes into contact with the post that they have the potential for *liking*. The degree to which an agent likes a particular *brand post* is calculated as:

$$L_{ij} = wspSP_{ij} + \sum_{k=1}^{n} w_k * P_{jk} \tag{6.3}$$

where L_{ij} is the utility of agent *i* for *brand post* j, wsp_i represents the agent's sensitivity to social pressure, SP_{ij} represents the social pressure experienced by agent *i* to like *brand post* j , and w_{ijk} is the agent's preference weight for *brand post j*'s attribute *k* P_{jk}

The equation above shows the agent's additive utility function for *liking* a post, in which each of the attributes of the post and the social pressure are multiplied by a preference weight to give an overall level of preference for the post.

Each agent has a preference for each of the attributes, drawn from a random distribution between 0 and 1, and a threshold. If L_{ij} exceeds the threshold then the agent takes the action *like*.

Having defined the theory and model that will be used in this application, the next section gives details of the data that will form the basis of the evaluation.

6.6 Empirical Data

This section contains a description of the empirical data to be used in this evaluation. The data was extracted from the social network Application Protocol Interface for the UK brand page of a major international coffee shop chain, covering one year's posts made by the brand, and coded according to their characteristics. In the particular social network under study, the posts contain a combination of text and photos. A typical post consists of a short statement, with a photo of a product. In an automated system it is more difficult to evaluate qualities such as vividness and value, but an algorithm was used to parse salient elements of the post into separate variables. Table 6.1 shows the definitions of the variables that were derived from the information provided by the API. In addition data was extracted from three other APIs to control for exogenous environmental factors. Finally, the degree of social pressure experienced by each agent is calculated for each *brand post* event, shown in Table 6.2. The choice of external data fed into the model is based on a combination of the data that was publicly available and a review of existing studies which have looked at brand posts in terms of the characteristics of their content, including their entertainment value, vividness, interactivity and informational quality, as well as control variables such as day of week and message length [46]. Each *brand post* record carries the date of when it was posted, but not a time series of the dates on which the *like* events occurred.

6.6.1 Model Calibration

This section describes the method used to calibrate the agent based model, taking into account the theory and data described above. The model was calibrated using the broad approach outlined in Section 5.3.2. When it was first presented in Section 5.3.2 the generic search process was conceptualised as a hierarchy entailing two related types of search — theory search and parameter search nested within it. This algorithm differs from that approach in that the two types of search are effectively combined and operate in parallel. In other words, parameters relating to the preferences are not calibrated separately as a sub-loop within each model iteration, but are searched for simultaneously with the theory elements as part of the overall procedure.

Within each iteration, the operations of each agent are processed in sequence.

- *For each agent:*

    1. *Perceive social pressure*
    2. *Perceive brand post attributes*
    3. *Evaluate utility for liking*

4. *Act*
5. *Update Memory*

- *Next agent*

The order in which agents are processed is randomised in each iteration.

For each of the variables discussed above, a parameter range was defined that reflected the constraints imposed by the relationship rules defined as part of the theory specification and detailed in 6.3.2. So for example if a relationship is expected to be positive or zero, only positive or zero parameter candidates are defined in the parameter range for that variable. The search procedure differs to that detailed in Figure 3.3 in that the search also includes elements of the theory, rather than just the parameters that are in the model. In this case the theory elements are the three possible network configurations outlined in Section 6.5.1. The genetic algorithm used to calibrate the model is shown in Figure 6.4.

The calibration process works against the parameters in the agent's utility function described in Section 6.5.5 and the choice of possible networks outlined in Section 6.5.1. Since there are sources of stochasticity in the model, and the genetic algorithm itself is stochastic, the calibration process was repeated 30 times to verify that it converged on stable results. The details of the results across the 30 iterations are shown in Figure 9.

This section described the process used to calibrate the model. The next section presents the results of the calibration exercise.

6.7 Results

In this section the results of the calibration process described in the previous section are presented. The outcome of the model calibration exercise can be considered in terms of three groups of parameters:

- *Agent sensitivity to exogenous factors* — the factors that were measured as being positive in the model were *product mentions* and *weekdays*. The presence of question marks, the length of a post and *posted at weekend* were negative. The other exogenous factors had a coefficient of zero. Because for some variables the range of parameters explored was constrained by the relationship rules described in Section 6.3.2, parameters that may have otherwise violated the expected sign of the relationship returned a parameter of zero.

- *Agent sensitivity to endogenous factors* — the parameter for *social pressure* was positive, meaning that the *like* actions of other agents in the social network were an influence on the probability of an agent deploying a *like* action themselves.

- *Theory features* — the network with the lowest concentration of longer paths between cliques was identified as being the most probable network - essentially the candidate that was associated with the best fit to the data.

Figure 6.4: The search algorithm used to find the macro level model and theory parameters

1. *Create initial population of* 150 *agent based models with each bit in the chromosome representing a parameter value for the parameters* wsp, *the* w_j *parameters associated with the* PB_i *in equation 6.3, and also the three candidate network specifications in 6.5.1. For parameters* wsp *and* w_j *the initial parameter values were drawn at random from the permissible (theory compliant) parameter range defined by the criteria set out in Section 6.3.2. For the three candidate network specifications the initial parameters were drawn at random.*
2. *Repeat until termination criteria met*
   (a) *For each agent based model*
      i. *Evaluate fit of the model using Equation* 3.4
   (b) *Score agent based models based on level of fit and theory score*
   (c) *Create new population of models. Reproduce models proportionally to overall score*
   (d) *Random crossover on top* 20 *equation pairs with two parents and probability* 50%
   (e) *Random mutation on top* 30 *of the new equation population with probability* 5% *in each bit*
   (f) *Check for termination criteria — whether iterations exceed* 200
   (g) *Loop, if not terminating*
3. *Exit procedure if terminating*

6.7.1 Comparing the results with other methods

To benchmark the level of fit, a linear regression model was also run using the same data, of the form:

$$\hat{Y} = \hat{\beta}_0 + \sum_{n=1}^{N} \hat{\beta}_i x_n \tag{6.4}$$

where $\hat{Y}$ is the predicted number of *likes* for post i, the $\hat{\beta}$s reflect the weights estimated for each of the characteristics x_n described in Tables 6.1 and 6.2.

The linear regression model above calculates the weightings required on the input variables to minimise the difference between $\hat{Y}$, its prediction of the aggregated number of *likes* that a combination of attributes would generate, and the actual number of likes that a particular *brand post* achieves.

The regression model was parameterised using the same method as that described in Figure 5.3.2.1 — an automated variable selection process using a genetic algorithm to search for a combination of variables that exhibit the permissible sign relationships. The same constraints were applied in terms of permitted signs as those applied through the relationship rules in the agent based approach described above.

The agent based model approach achieved a reasonable level of fit — an R-Squared of 63%, while the linear regression model achieved an R—Squared of 56%. Both levels of fit are calculated using using equation 3.4. The difference between the degree of fit achieved by this linear formulation and that achieved by the ABM approach could be interpreted to be the incremental contribution of the sharing impact. Although the level of fit in both cases suggests that there are additional factors that should be considered in order to explain the total level of *likes* for each post, the ABM's capability for accommodating emergent features appears to improve the model's explanatory power.

6.7.2 The Impact of Agent Processing Order on the Results

As mentioned in Section 6.5, agent operations in the model are processed sequentially. This means that within any particular iteration the behaviours of the agents processed earlier in the queue may begin to influence the behaviour of agents processed later in the queue within the same iteration, through the social pressure mechanism described in Section 6.5.1. This is distinct from the approach of equation based methods like the first order differential equation described in Section 2.6.4.1, in that social diffusion can be represented within a particular event, rather than just across time periods.

Because the data is available at aggregate level rather than for each individual, and because each record refers to a brand post event rather than a full history of likes for the post over time, there are limitations to the degree to which the social pressure element can be interpreted. A consideration during the course of the analysis was that when modelling social diffusion within a single iteration the order of execution may have an impact on the outcome of the simulation in any particular iteration. For example, if the order of execution allocates agents with a low level of preference for the attributes of the target entity earlier in the execution queue this is less likely to lead to a high degree of social sharing in that iteration than if the order of execution was reversed.

To test the impact of agent execution order on the social pressure component of the model, a simple simulation was run with a set of 1000 (N) agents. Each agent was given a preference for a nominal product ($prodpref$), and also a sensitivity to social pressure (SP), with each drawn independently from a random uniform distribution between 0 and 1, such that there was no correlation at a total level between product preference and sensitivity to social pressure. The agents were placed in a social network identical to that described in Section 6.5.1, with access to information about the product purchases of the agents to whom they were linked. The agents were given a simple utility function, and if their utility function exceeded a threshold they purchased the product, as below:

$$Purchase_{it} = ((0.1 * SP_{it} + 1 * prodpref_i) > 0.5) \tag{6.5}$$

where $Purchase_{it}$ is an agent i's purchase action in iteration t, SP_{it} reflects the number of neighbouring agents who have already purchased in iteration t and $prod_pref_i$ reflects the agent's preference for the product.

The sum of total purchases in each iteration is calculated as being the sum of agents for whom utility exceeds the threshold:

$$Sumofpurchases_t = \sum_{i=1}^{N} Purchase_{it} \tag{6.6}$$

where N is the total number of agents in the simulation and $Purchase_{it}$ is an agent i's purchase action in iteration t.

In other words there are two factors affecting the agent's decision, pressure from his neighbours, SP_i, and his own preference $prodpref$. To test the impact of the order of execution on the outcome of the simulation in a particular iteration — the total number of purchases $Sumofpurchases_t$ — three experiments were run:

1. One iteration of the simulation with the agents executing their behaviour in descending order of $prodpref$, in other words the agents with the highest level of $prodpref$ first.

2. One iteration of the simulation with the agents executing their behaviour in ascending order of $prodpref$, in other words the agents with the lowest $prodpref$ first.

3. One iteration of the simulation with the agents executing their behaviour in a random order.

Figure 6.5 shows the results of one run of the simulation. It is clear that the growth in aggregate purchases within the iteration and the total number of purchases attained vary substantially across the different cases:

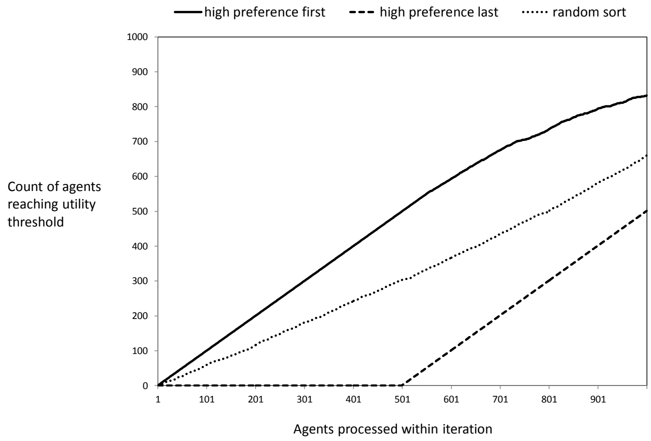

Figure 6.5: Purchases achieved in one iteration with varying orders of execution

1. Under the first case, with agents executing in descending order of product preference, aggregate purchases grow quickly and reach a sum of 832 within the iteration.

2. In the second case, with agents executing in ascending order of product preference, no purchases are made until the agents with a $prodpref$ of greater than 0.5 are reached in the execution queue, and total purchases, $Sumofpurchases_t$, sum to 500 — essentially all of the agents with a $prodpref >$ 0.5.

3. In the final case, with agents sorted randomly, social pressure (SP) also has an effect, with the sum of 660 agents making purchases at the end of the iteration.

In summary, until the agents with a preference for the product begin making purchases, no social pressure is created to influence the agents with a lower product preference. If the lower product preference agents have already taken their turn in the execution queue then the social pressure effect created by the higher product preference agents comes too late to influence them and the overall level of sales achieved is lower.

6.7.3 Recommended Actions

This section describes the process used to derive implications from the model and presents some of the recommendations. In Section 2.6.2 the necessity for marketing managers to make decisions regarding policies in complex and dynamic environments was discussed, and this necessity in turn formed the requirement outlined in Section 5.4 for an automated model selection process to be able to propose actions based on the theory that it has learnt. In this implementation a set of IF THEN rules is applied to the selected theory to produce a set of recommendations about how brand posts for the particular advertiser could be adjusted to increase the level of *likes* acquired by future posts.

The theory specification described in Section 6.3 laid out a number of possible preferences that the agents could adopt. Recommendations rules were then developed in the form of 'If post feature X is a positive influence then do more of it', 'If post feature X is a negative influence then do less of it' to apply to the factors laid out in Table 6.1. In addition rules were developed to apply to the environmental factors set out in Table 6.2. These rules were applied to the set of preferences that were discovered, through the model calibration process, to be active in the agent's utility calculation. As discussed in Section 5.4, the factors affecting the level of certainty for a rule antecedent or consequent are likely to be specific to the particular domain. In this application, the rules were ranked based on a combination of the magnitude of the effect of each feature on the number of *likes*. Of these, the top scoring rules were:

Factor	Parameter Value	Interpretation
Days since last post	0	Post more frequently
Length of post	-	Use shorter posts
Product Mentioned	+	Include product mentions more frequently
Posted at Weekend	-	Avoid posting at the weekend
Posted on Weekday 2	+	Post more on weekday 2
Posted on Weekday 3	+	Post more on weekday 3

The recommendations provided by the system proposed are useful in the sense that they provide the kind of insight that could be used to guide marketing policy decisions. In Chapter 8 some additional methods for assessing the practical utility of automatically derived recommendations are considered.

6.8 Conclusions

In this chapter an application of the proposed theory selection system to an empirical dataset was described. The agent-centric theory developed in Chapter 4 was parameterised to reflect the elements and potential relationships in an online social networking domain. The elements of the theory representation were mapped to the elements of an agent based model. The best model and theory were then searched for using the method described in Chapter 5 — working on a macro level empirical data set. Elements of the theory were selected from multiple parts of the ontology — endogenous features, exogenous factors and network structure. Recommendations were derived, based on the model outcomes, using the rule based approach put forward in Section 5.4.

In terms of the overall research objective — developing and evaluating a solution to the problem of selecting an agent based model specification from a set of possible specifications — the conclusions that can be drawn from this chapter can be split into three broad areas:

- *Evaluating the proposed framework* — The approach proved to be successful in selecting a model and theory from the wide range of parameter values and variable combinations that formed candidates.

Given that the model was successfully selected, some consideration should be given to whether or not it is a *good* model. In the applied evaluation in Chapter 3 it was possible to determine whether the correct model was being chosen because the data was artificially generated. When working with empirical data there are no absolute criteria for assessing whether a model is *right*, but there are some criteria that are commonly applied.

As discussed in Section 5.2, model accuracy is a measure of the extent to which the *mDGP* creates output which is similar to the target: *rwDGP*, and Grune-Yanoff argues that if a model cannot reproduce the target state, there is no actual explanandum of it [85]. The agent based model achieved a reasonable level of fit to the empirical data, with an R-Squared of 63%, and exceeded the performance of the linear regression model against the same dataset. There is not an accepted threshold in the ABM literature for the degree of fit above which a model can be considered to have reproduced the target state. In Section 8.3.2 these results are discussed and evaluated more fully, and considered in the context of the results from the next chapter, Chapter 7. Apart from the level of fit to the data, the model can also be assessed based on the relevance of the recommendations which it implied for a marketing management environment. Some methods that could be used for evaluating relevance are discussed in Chapter 8.

- *Broader methodological issues* — In terms of its methodological implications for agent based modelling, the study conducted in this chapter suggests that the value of the agent based method relative to approaches which do not take account of social diffusion is particularly likely to be felt where the target process includes interactions between constituent entities, as is often believed to be the case in processes in the marketing domain. It may be worthwhile investigating the difference between emergent effects which are consequences of certain thresholds being met — which may be due to other factors outside of the model — and effects which result from endogenous factors, for example the specific constituents of an event, in our application a *brand post*. If specific event elements are significant then the total potential emergent feature may be limited by the preference distributions of the entities operating in the target environment.

  As discussed in Section 2.3.14, random influences are commonly found in agent based models, and the order of execution issue identified in Section 6.7 introduces noise into the process which poses a threat to the model's ability to select the best model. This threat can be reduced with multiple runs of the model to calculate a stable average [32].

  The results from the exercise suggested that there was a combination of endogenous and exogenous factors that were significant in the target process. The ability to represent these features separately means that with more granular data, more complex scenarios could be explored in which order of exposure is considered. For example, in some target processes it could be the case that the early *likers* are individuals with a high degree of preference for the properties of the *brand post* and later *likers* are individuals with a high degree of sensitivity to social pressure. With time series data, this kind of analysis might become possible.

As well as the immediate impact of direct social interaction between agents, in the real world social network there is a *global* indicator which gives a count of the total number of *like* events for a *brand post* by all individuals, regardless of whether or not they are in the individual's social network. Because the data was only available at total post level, the global indicator was not represented here. This global indicator could be considered to be similar to the factor which Conte [41] describes — emerging through agent interactions and in turn having an impact on the behaviour of the agents in the model which is distinct from the impact of the individual interactions. This is arguably a form of downward causation of the type discussed in Section 4.2.1 and with more detailed data availability — perhaps at daily level — it could be incorporated into the simulation as a second type of social pressure. The *global* like count could be determined endogenously in the model and incremented with each additional *like* and made visible to each of the agents whose operations had not yet been processed, meaning that social pressure from outside of their immediate network was taken into account.

- *Applied marketing research* — From the perspective of applied marketing research, deploying the approach in a social marketing environment builds on the concepts put forward by Delre *et al.* [47] and North *et al.* [154], discussed in Section 2.6.4.2 regarding modelling of social diffusion. It also builds on the practical applications of Cha et al. [35], Sun et al. [190] and Yu and Fei [210] by applying an aggregate level model to an online social diffusion process and linking the drivers of diffusion to specific elements of online content.

  While existing work such as that by Watts and Dodds [202] considers the role of early adopters — agents who are activated after being influenced by just one neighbour — in triggering cascades, the approach used here adds the potential to consider the role of other preferences in creating such cascades, assuming that data is available at a more granular time series level that includes information on individual preferences.

The data used in this chapter was retrieved at an aggregated, macro level and the methodology used exploited ABM's ability to represent social diffusion, but not its capabilities with regard to representing heterogenous micro-level entities. In the next chapter, Chapter 7, a micro-level data set is used to evaluate the utility of the approach against more granular data — where information is collected about the behaviours of individual households.

Chapter 7

Theory Selection with Micro-Level Empirical Data

7.1 Introduction

In this chapter an implementation of the modelling framework described in Chapters 4 and 5 is evaluated against an empirical, micro-level data set. The aim of the implementation is to evaluate the proposed solution to the problem of selecting an agent based model specification from a set of possible candidates using individual level data. This application broadens the range of agent based theory selection problems that the proposed framework is evaluated against and differs from the evaluation described earlier in Chapter 6 in three key respects. Firstly, it uses optimisation methods from existing statistical modelling methods to calibrate the models rather than a genetic algorithm. Secondly, the empirical data used is observed and recorded at individual level rather than at an aggregate level, meaning that agent models can be developed at a micro-level and therefore exhibit heterogeneous characteristics. Finally, empirical fit is evaluated based on event level predictive accuracy, rather than over time as in earlier chapters.

In Section 7.2 a brief overview of micro level data is presented, alongside a review of existing modelling methodologies that have been applied to it in different domains. Then in Section 7.3 the specific elements of the theory structure used in this analysis are outlined. In Section 7.5 the background to the specific data set that is under study in this chapter is described. this data contains information that has been retrieved and stored at a micro level and represents the outcomes of a real world data generating process. In Section 7.6 the statistical methods used to calibrate the agent's behavioural models are presented. In Section 7.7 the results of the modelling exercise are presented and discussed. A comparison of a model performance with and without social interaction is also presented. Then in Section 7.8, a series of simulations is run based on the agent models to test the impact of policy decisions on emergent behaviour in the simulated environment.

7.2 Background to Modelling with Micro Level Data

This section presents an overview of micro-level data, the various methods that have been used to analyse it, and the potential to enhance these methods using concepts from agent based modelling.

Data availability has often been a problem for developing detailed predictive models at individual agent level. The data available to social scientists has historically been limited, generally self-reported and often static — collected at a single point in time. But new technologies increasingly offer real time records of human behaviour and interactions over extended time periods, sometimes providing information about the structure and content of relationships as well [124]. The increase in data availability has been brought about by developments including the digitisation of information that was previously held in analogue form, and the proliferation of digital devices that provide real-time reporting [114]. This includes information on credit card transactions, movements of products with radio-frequency identification (RFIDs), and online product searches and purchases [114]. Structured records, repeatedly collected over time for particular individuals or entities, form a kind of longitudinal dataset which can be used for a range of modelling purposes [84], and open up the possibility of parameterising ABMs empirically at individual level. Mitchell envisages that some of the areas where this kind of data can be applied are in reducing traffic congestion and pollution, limiting the spread of disease and optimising the performance of public resources such as transport and emergency services [146].

While many individual based methods already share characteristics with ABM, such as heterogeneous properties and preferences, and the ability to represent bounded rationality, combining these existing methods with ABM's support for interactions opens up the possibility of modelling the behaviour of all the people within a group of individuals while taking into account the structure of their social ties and the adaptive behaviour that might follow from these.

7.3 Theory Specification

In this section the theory representation used in this analysis is described. The overall structure of the theory representation, based on the framework laid out in Section 4.2.2, is illustrated in Figure 7.1. The theory structure contains product features and advertising as exogenous variables, characteristics of the agents — in this case properties and behaviours. Finally the specification contains an emergent feature — *product buzz*, which reflects the level of discussion between the agents of the products in the market. Sections 7.3.1 to 7.4.1 below describe the components in more detail.

7.3.1 Exogenous Factors

This section outlines the exogenous factors that have an influence in this simulation. A key element of the theory specification laid out in Section 4.2.2 was the role of exogenous factors — facts that are introduced from outside of the simulation. Table 7.1 shows the exogenous factors that are relevant in this application — product characteristics that are introduced into the model by manufacturers and perceived by the agents.

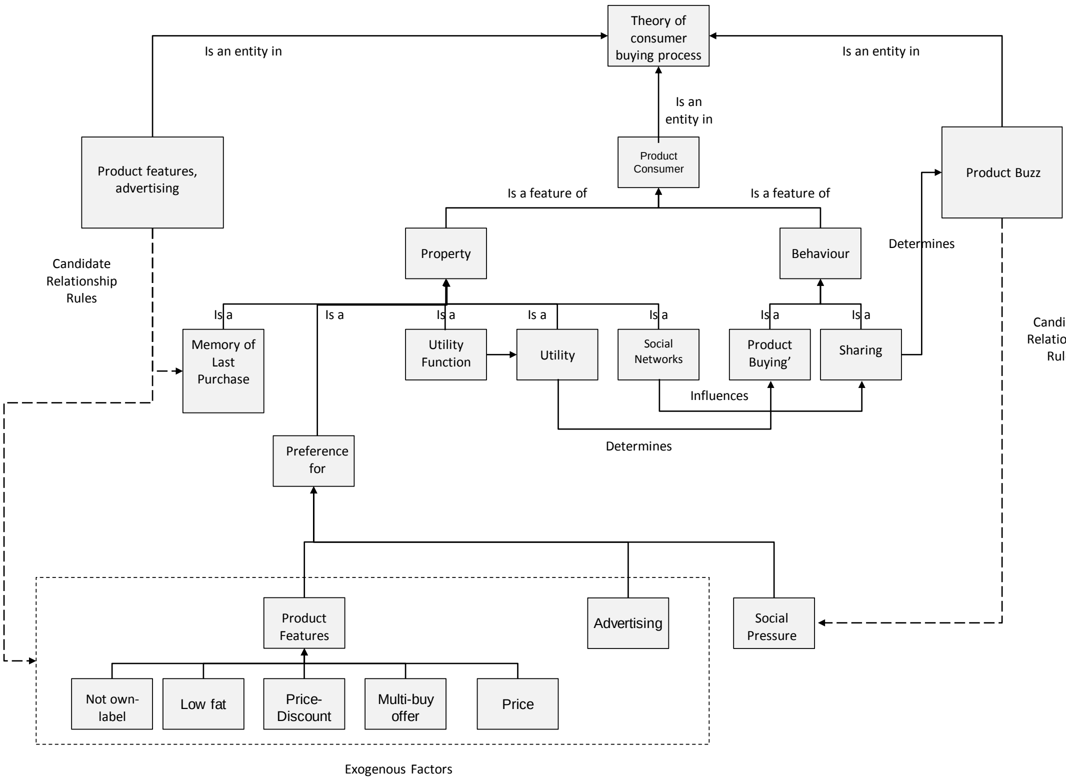

Figure 7.1: Definition of the theory representation used in this chapter

Variable Name	Variable Label	Variable Description
exWeight	Weight	The weight of the product, expressed in grams. If the product is part of a multi-buy offer the weight refers to the standard pack weight multiplied by the total number of units offered as part of the multi-buy
exPrice	Price	The selling price of the product per gram, taking into account any price discounts or multi-buys
exNotOwnLabel	Not-own-label	A dummy variable that indicates whether or not the product is a supermarket own-label
exAdvertising	Advertising	The level of television advertising for the product, expressed in GRPs (Gross Rating Points) and decayed by 50 percent per month
exPriceDiscount	Price Discount	The percentage price discount — (standard price - discounted price)/standard price
exMultiBuyOffer	Multi-buy Offer	The discount level of a multi buy offer, for example a two for one is expressed as a fifty percent offer
exLowFat	Low-fat	A dummy variable reflecting whether or not the product claims to have *low fat content*

Table 7.1: Exogenous variables — characteristics of products in the yellow fat market

7.4 Mapping Rules

Figure 7.2 shows the mapping rules used to connect the theoretical specification to the models in this section. The mapping rules described here differ in a number of respects to those laid out in Chapter 6. The key differences are that only one social network topology is being tested and three possible agent states (detailed in Section 7.4.0.2) are defined. Apart from these differences the other parts of the rule are common to both chapters, in that the first group of items, starting with ex, represent the exogenous variables that are active candidates in the simulation, the second group, starting with $pref$ represent the agent preferences that these correspond to, and the third group represents constraints on the parameter values that are acceptable.

7.4.0.1 Social Pressure

This section describes the social interactions between agents in the model. Each agent in the modelling process is situated in a social network. As in Section 3.5, a small world network was created using the methodology proposed by Watts and Strogatz [202] — assigning each of the individuals to a node on a regular network, then randomly rewired the links until the network showed the properties of a small world — a mixture of short paths connecting most of the individuals within each clique, and longer paths connecting the cliques. The structure of the small world network means that trends are more likely to

Figure 7.2: The mapping rules used to connect the agent based model with the theory parameters

([exWeight, exPrice, exNotOwnLabel, exAdvertising, exPriceDiscount, exMultiBuyOffer , exLowFat],[prefWeight, prefPrice, prefNotOwnLabel, prefAdvertising, prefPriceDiscount, prefMultiBuyOffer , prefLowFat],$[-1:1,-1:0,-1:1,,0:1,,-1:0,,0:1,,-1:1]$,*[utility-additive],[social-network-1],[neighbour-pressure],[full-period],[state-1, state-2, state-3])*

emerge locally amongst the immediate group, but that emerging trends amongst the local group can be propagated more widely through the longer paths. The agents experience social pressure as:

$$SP_{ij} = \frac{\hat{N}_j}{N} \tag{7.1}$$

where SP_{ij} is the social pressure felt by agent *i* to purchase product *j*, $\hat{N}_j$ is the number of agents with whom agent *i* is linked who are currently expected to consume product *j*, and *N* is the total number of agents to whom agent *i* is linked.

In other words, as the proportion of the agents who are linked to agent i and the model predicts will purchase the product increases, the social pressure on *i* also increases. This SP_{ij} for each of the j candidate choices is considered in the discrete choice model (described in Section 7.6 below) as a candidate factor for each i individual. Because the choice model is estimated separately for each individual there is the potential for a heterogeneous response to social pressure within any particular group. There is a possibility that some individuals may choose to avoid group trends, and so may have a negative preference for *social pressure*.

7.4.0.2 States

This section describes the different states that agents in the simulation may occupy. In Section 4.2.2.1 it was noted that in some cases a state may be directly observable, and in others may be observable only through another behaviour. In this application, three discrete underlying usage disposition states are defined that each individual might occupy: Low, Medium, and High. These unobserved states are mapped to levels of observed behaviour as follows:

- State 1 — *Low Disposition* = 0 or 1 purchases per month
- State 2 — *Medium Disposition* = 1 or 2 purchases per month
- State 3 — *High Disposition* = 2 or more purchases per month

The initial transition probabilities (presented in Section 7.6) are estimated from observation of the natural switching in the data and allowed for a number of covariates, representing impacts from the environment, to affect the likelihood of switching between states. The covariates were the key seasonal periods of Christmas and Easter. The mechanics of state switching are described in more detail in Section 7.6.

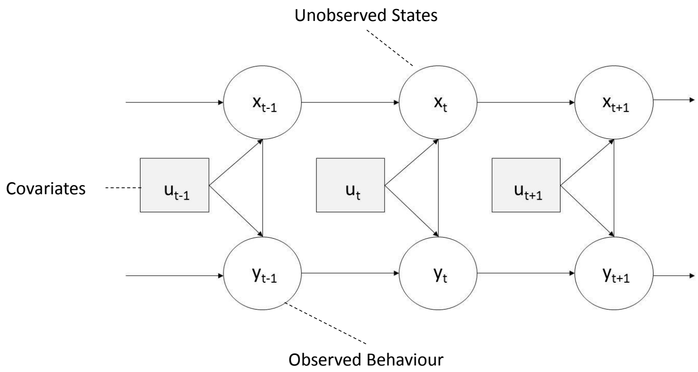

Figure 7.3: Input output hidden Markov model

7.4.0.3 Actions

In this application, the actions that agents are able to take are to either make no purchase in a particular time period, or to purchase one of a number of possible products for sale. The features of the available products are detailed in Table 7.1.

7.4.0.4 Memory

This section presents the memory functions of the agents in this evaluation. Agent memory is comprised of the *Ever bought before* variable decayed by a parameter λ per purchase occasion to reflect how recently (in terms of purchase occasions) a product was last purchased. This forms a partial memory of the agent's previous purchases. The approach to representing the memory function is similar to that described in the aggregate application in Section 6.5.4, with the key difference being that each agent may have a different parameter.

$$EverBoughtBeforeMemory_{ijt} = BoughtInThisPeriod_{ij} + \lambda EverBoughtBeforeMemory_{ijt-1} \tag{7.2}$$

where $EverBoughtBeforeMemory_{ijt}$ is each agent i's memory in the current time period t of having purchased a particular brand j before, $BoughtInThisPeriod_{i}j$ is a binary variable reflecting whether they are buying that brand in this period and λ is the decay parameter which diminishes the memory of their previous purchase.

The equation above reflects the operations of the memory that each agent i has of having previously purchased the particular brand j before. As the number of occasions since his last purchase of that particular product increases, his memory of having bought the product before decreases by the factor λ.

7.4.1 Emergent Factors

In this section the factors that are endogenous to the model are explained. In Section 2.3.7.3, various methods for representing social interaction were discussed, some of which are based on social pressure and some on message transfer. The social interaction representation used here differs from the social pressure mechanism implemented in Chapter 6 in that it reflects discussion of the product by consumers. But in common with the approach used in Chapter 6, social pressure emerges from the interaction of the agents in the social network. After each time period, each agent discusses purchasing habits with each of the other individuals to which it has a direct link. In the model calibration phase, each agent collects data about the other agent's modelled preference, i.e. the action that maximises their expected utility based on their estimated preferences, rather than collecting information about the actual action taken. This is to avoid the effect of the real observed actions becoming unduly influential in the estimation phase, since real information will not be available in the validation and simulation phases. This modelled information about the actions taken by the agent's network forms an emergent social pressure effect.

7.5 Empirical Data

In this section an overview is presented of the empirical dataset against which the proposed research solution is evaluated. The dataset used in this chapter was created at a micro level by customers shopping in a particular retail food category who make a sequence of product choices over a number of purchase occasions, for example the sequence A, A, B, A, A, B might reflect the purchases of one customer in a market with two products, *A* and *B*. At a macro level, these purchases aggregate to form the sales patterns of individual brands and firms, with the impact of competitive market interventions by vendors, including discounts and price changes, competitive packaging etc. leading to changes in consumer behaviour which in turn leads to volatility in the aggregate series [180]. Some of the key features of the data are listed below:

- The data used for the analysis was collected by a panel of households who recorded their purchases using scanners, and initially comprised 2 million purchases in the yellow fats (butter, margarine) category over 3 years. Some of the households were eliminated from the data for having fewer than 5 transactions over the time period, and the scope of the data was narrowed to include only households who are primarily purchasers of margarine. The scope was narrowed to maintain the assumption of independence of irrelevant alternatives (IIA).

- For all of the constituent panel households, a unique product code is recorded for each purchase in the category, along with the date, retailer, offer and price paid. This is supplemented with detailed characteristics that are assigned to the product code, and additional information about the retailer and average price for the product. In addition, advertising data was appended using dates as a key.

- The data is indexed by a house ID which represents a unique scanner, and is the basis for the definition of an agent. There are several issues which complicate analysis of the data. Kahn *et al.* [104] note that aggregation of individual level behaviour to a household level can mask the decision making criteria of the individual actors in a household. This is likely to be more acute in product categories where household members have individual needs or tastes. In addition, the data does not contain information about the display features used in-store. Shaw and Merrick identify display as a major factor in promotions, and also in highlighting distribution [183].

- Environmental data — data reflecting processes that are not affected by agent behaviour in the model itself — is collected simultaneously with the observation of the agent's behaviour, and its role in the simulation is learnt indirectly through its impact on the state and action selections of the agents.

- Product distribution data for this category is not explicitly available but can be inferred from customer purchases, for most products by week and by store; but if a purchase was not made by any household for a particular product/ store/ week permutation, that product will be absent from the choice set of customers shopping in that store. Finally, the network of contacts between households, if any, is unknown.

- The dataset contains 150 calendar weeks, with most households making a purchase every 5-6 weeks. The first 110 weeks of data were used to estimate the model's coefficients, and the last 40 weeks form an out-of-sample validation period on which the predictive power of the model is tested [45].

7.6 Model Calibration

In this section the model calibration procedure used in this application is detailed. The model was initialised using the first 30 periods of the data, providing a stable history for the agent memory and social pressure inputs. The models are then run iteratively for each agent for each time period for the remaining 80 periods of the training data. In each period the modelled memories and social pressure variables are updated, meaning that if appropriate, interactions between agents in earlier periods can influence their decisions in later periods. Each agent converges towards a more stable behavioural model as additional data is added.

The remaining 40 periods are used to validate the models. The observed environmental data for the validation period are used as inputs, but the purchase memories of the individual agents and the actual purchases of other agents in the environment are estimated, so that no new real information is introduced into the system which may influence the predictions.

A hidden Markov model was used to estimate state transition probabilities for each individual agent, subject to a set of covariates, defined in Section 7.4.0.2. Once a set of possible states have been defined, hidden Markov models provide a method of learning state transition matrices — the probability with which agents will switch between states in any time period. The table below shows the structure of the transition matrix. Each cell represents the probability that a given agent will move from a given state at time *t-1* to a different state at time *t*.

	State at *t*: Low	State at *t*: Medium	State at *t*: High
State at *t*-1: Low	0.54	0.31	0.15
State at *t*-1: Medium	0.27	0.62	0.11
State at *t*-1: High	0.12	0.56	0.32

Unlike in a standard Markov model, where the state is directly observable and the state transition probabilities are the only parameters, hidden Markov models assume that the actual underlying state is not directly observable, but only the variables influenced by the state. Hidden Markov models can be extended to accept exogenous inputs that affect the underlying state transition probabilities, meaning that they can take into account the environmental conditions that may cause agents to switch between states. Figure 7.3 shows how the combination of exogenous variables u_t and previous states may affect the transition between the underlying and observable states, where x_t represent the sequence of unobserved states and y_t represents the sequence of observed states. A different model was estimated for each of the individuals for whom data was observed. This was to allow for the possibility of a heterogeneous response to the covariates and also to accommodate differences in the natural level of demand for the product between individuals.

The agent level action selection models are implemented using a multinomial logit discrete choice procedure [170], which is commonly used in choice scenarios where there are three or more options. The probability that an individual will choose brand j is calculated as:

$$P(y_i = j) = P_{ij} = \frac{exp(x_i'\beta_j)}{\sum_{k=0}^{j} exp(x_i'\beta_j)} \tag{7.3}$$

where y_i is a random variable that indicates the choice made, x_i is a vector of characteristics specific to the ith individual, and β_j is a vector of coefficients specific to the jth alternative.

In other words, an estimate is made for each household i is about the set of weightings β_j that they give to the particular characteristics of option j. The probability that household i will choose the particular option j is calculated as the combined score of the weightings that they assign to particular attributes and the attributes that the product possesses. The household is considered to choose the product that they have the maximum probability of buying.

Discrete choice models were used to learn the underlying agent preferences that influence an agent's choice of action. In a discrete choice model a decision maker uses a decision rule to distinguish between a set of alternatives, each of which has attributes, for which the decision maker has preferences [21]. The assumption that preferences are transitive is used, and IIA (independence of irrelevant alternatives) is also assumed — i.e. that the ratio of choice probabilities of any two products does not change when the consideration set changes as long as both of those products are in that set [170]. Discrete variables can be either binary (possessing only two possible values or outcomes) or multinomial (possessing more than two possible discrete outcomes).

Based on the domain theory outlined above, the set of factors described in Section 7.3.1 was derived and tested in the model selection process with appropriate restrictions expressed by the relationship rules. The thirteen candidate factors are split into those that are observed from the agent's environment and those that are observed or derived from the agent behaviour (own and others).

In earlier chapters, genetic algorithms were recommended and used to search the more complex search spaces created by the interactions between theory elements and agent parameters in integrated agent models. Because the logistic function defined in 7.3 and used in the discrete choice model is commonly employed for similar purposes, for example in [170], and creates a well understood search space, and because the agent parameters are estimated individually rather than as part of a total system, the individual agent level discrete choice model was calibrated using the faster quasi-Newton search method in *Proc MDC* in *SAS version 9.1*, SAS Institute Inc. Quasi-Newton methods minimize a quadratic function in n iterations, where n is the number of variables. Constraints to the parameters, passed by the rules laid out in Section 7.4 were applied using the built in *bounds* statement. As above, because the hidden Markov model was calibrated at individual level, the faster Nelder—Mead method was used, which was the default optimisation algorithm in version 1.3 of the *msm* package in *R*.

7.7 Modelling Results

The preceding sections outline an approach to parameterising individual agents using a range of existing learning methods. In this section the approach was applied in order to test whether this ABM framework can produce more accurate forecasts than non-ABM methods by running the model with and without the features of state transition and communication that distinguish the ABM approach, and evaluate the usefulness of the simulations that the learnt models can establish. This is done by creating an individual level model of consumer behaviour using multinomial discrete choice estimation and hidden Markov models, then testing it against a second model in which agent behaviour includes the ability to communicate within social groups. The behavioural model for each agent, consisting of state transition probabilities, preferences, and an action selection algorithm is individually calibrated using data about the individual's observed behaviour, the observed behaviour of other individuals in the same environment, and data observed about the environment. From a qualitative viewpoint, the implications of the simulated scenarios that this kind of approach produces are reviewed.

Figure 7.2 shows the prediction accuracy at brand level that it was possible to achieve with and without a networked model. The numbers reflect the degree of fit to the empirical data, expressed as the proportion of cases for which the model predicted the correct purchase. The results are also split into the level of accuracy achieved in the calibration phase and the validation phase. Overall the degree of accuracy seems reasonable in the estimation period, but the fact that a number of the brand choices could not be predicted

State	With Social Pressure Mechanism	Without Social Pressure Mechanism
Calibration Period	66%	63%
Validation Period	53%	52%

Table 7.2: Prediction accuracy of models with and without social pressure component

	Social Pressure	Last Purchase	Multibuy	TPR	Advertising	Price	Weight	Non-label	Low Fat
Social Pressure	1								
Last Purchase	-0.1	1							
Multibuy	0.0	.34	1						
TPR	-0.01	0.37	0.76	1					
Advertising	0.00	0.00	0.00	0.00	1				
Price	-0.05	-0.04	0.01	0.01	0.00	1			
Weight	0.00	0.00	0.00	0.00	0.94	0.02	1		
Non-label	-0.39	0.13	-0.54	-0.54	0.00	0.00	0.00	1	
Low Fat	-0.09	0.16	-0.13	-0.01	0.00	0.03	0.00	0.03	1

Table 7.3: Correlation matrix of learnt preference coefficients across the agent population

by the model shows that there is still some unexplained variation in the purchase process. The model with social pressure included performs better than that without in both the calibration and the validation phases. In the calibration phase the performance is somewhat better, whilst in the validation phase the difference between the two is more limited.

7.8 Simulation

This section gives the details of a number of simulations that were carried out to test the implications of different marketing policy interventions. As discussed in Section 7.6, the calibrated models provide a heterogeneous set of preferences and state transitions for each agent. Figure 7.3 shows the preference correlations across the population of agents. The data is calculated by stacking the coefficients for the individual agents, then calculating the correlation across the coefficients for all of the agents simultaneously - for example there is a 76% correlation between preference for *multi-buy promotions* and preference for *TPR* (Temporary Price Reduction) offers - suggesting a high degree of cross-holding of these two preferences in the agent population. Although the majority of preferences for *social pressure* are positive, the results indicate that some agents prefer to be individualistic, showing a negative preference for other agent's behaviour. Figure 7.3 shows that sensitivity to social pressure is reasonably highly correlated with a rejection of *own-label* supermarket brands, i.e. a preference for branded products. In general agents who prefer promotions of either kind are less likely to prefer *own-label* supermarket products. The heterogeneous estimation approach results in a joint distribution of parameters, which in aggregate models would need to be assumed and imposed, and can be used to simulate the impact of future policy changes on a brand by brand basis.

To test the implications of these policy changes, a simulation environment was established using the agent models that resulted from the calibration process and an artificial time series data set. The artificial dataset reflects all of the properties in the original training dataset, described in Section 7.5, and any of the properties can be modified to test the response of the agents, given the parameters learnt from the initial calibration process. A number of scenarios were tested to assess the impact of policy changes:

In order to test the impact of marketing instruments on switching behaviour:

1. *A significant price discount by Brand C was introduced into the simulation at time period 10.*

2. *The percentage price discount — defined as (standard price - discounted price)/standard price — applied to Brand C was 50%.*

3. *The simulation was run and the switching behaviours from other brands and the uplift in Brand C sales were observed.*

Figure 7.4 shows the market share that each of the top selling brands achieves before and during the promotion. Because of the impact of agent memory (defined in Section 7.4.0.4) — reflecting either habit or brand loyalty, and the level of match between some brand's characteristics and the preference sets of their customers, some brands lose very few sales during the simulated promotion, whereas others lose up to 55 percent. For example *Brand F* loses very little, while *Brand E* loses half of its sales.

In order to test the long-term impact of a promotion on the sales of the promoting brand and its competitors, a simulation was also run representing a hypothetical discount on *Brand A*. A significant price discount by Brand A was introduced into the simulation at time period 9. The percentage price discount applied to *Brand A* was 60%. As before, the simulation was run and the switching behaviours from other brands and the changes in share for the different brands were observed over time. Figure 7.5 shows market share for each brand returning to its original level around five weeks after the price discount initiative for *Brand A*. Additional experiments were conducted, increasing the level of price discount on *Brand A* to 70% and then 80%. As with the initial price discount, market shares returned to their equilibrium level five to six weeks after the promotion was run.

The simulated market appeared to be too stable to create any kind of take-off momentum with realistic levels of interventions. This could be in part due to the distribution of preferences across product features, which are catered for by the various brands and sub-brands operating. Since, as Equation 7.3 shows, each

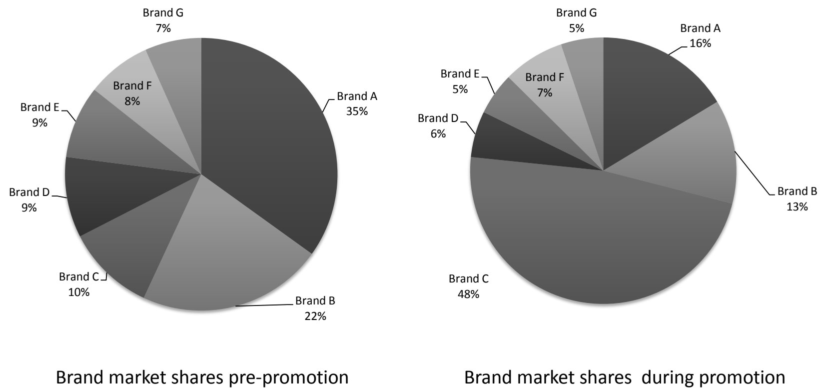

Figure 7.4: Market shares for biggest selling brands during and after promotions

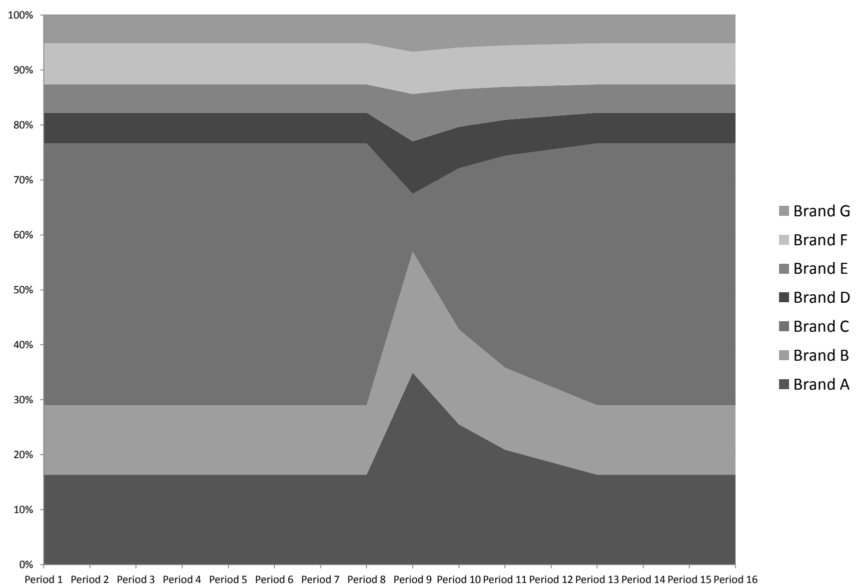

Figure 7.5: Market shares for biggest selling brands subject to promotional interventions

agent's utility for a particular brand in the simulation is based on a combination of βs and independent variables, the choice between brands is based on the balance of these effects within and across brands. If the preference for a particular brand or product feature are strong enough they will have the effect of limiting any migration to other brands within the same product-feature group. The social pressure variable (in Equation 7.1) or its β needs to have sufficient magnitude to overwhelm the other effects in the utility evaluation. As discussed above, because the agents have heterogeneous preferences there is a level of correlation between social pressure and certain preferences which makes some products less likely to be responsive to social pressure than others. In this case sensitivity to social pressure was negatively correlated with own-label products, suggesting that a preference for these products will be difficult to disturb using social pressure. In a market where all of the products have equivalent features and there is limited brand preference it may be easier to achieve take-off. Finally, the threshold for an emergent take-off may be higher than could be achieved by varying price and promotional activity within a reasonable range.

In addition the impact of the memory parameter means that agents who have repeatedly bought the same product or set of products are unlikely to switch, even if there is a strong price discount in a competing brand. In principle, the effect of social pressure on choice will also create a memory effect, since the effect of the dynamic on $EverBoughtBeforeMemory$, defined in Section 7.4.0.4, will be influenced by the process in Section 7.4.0.1 and shape future memories. Again, within a reasonable range of interventions this could not be achieved in the simulation at a level which shifted agents perpetually from their previous habitual brand range.

7.9 Conclusion

This chapter presented an application of the theory selection system proposed in Chapters 4 and 5 to an empirical dataset in which the data was gathered at a micro level. The theory design was customised to represent the factors that were relevant to the specific domain, and the models that were initialised were developed to reflect the decisions and stimuli relevant to individuals shopping for a product in a competitive market.

The objective of this chapter was to evaluate the proposed solution to the problem of selecting an agent based model from a range of candidates using an implementation of the solution proposed in Chapters 4 and 5. The evaluation approach used optimisation methods from existing statistical modelling methods to calibrate the models rather than a genetic algorithm, and to develop models at a micro-level that exhibit heterogeneous characteristics. The conclusions that can be drawn from this chapter can be split into three broad areas:

- *Evaluating the proposed framework* — Models and their theoretical implications were successfully selected using the framework put forward in Chapters 4 and 5. The resulting models derived from the process are useful in understanding the internal dynamics of the market and although this particular market appears to be relatively stable, in terms of its resistance to destabilising interventions, some domains may exhibit more interactions and emergent effects than are present in this environment. The three primary causes of stability in this simulation are product preferences, habit — the agent's previous purchase history, and social pressure. As discussed above, habit and social pressure interact through the effect of prior social pressure on prior purchases, which is then carried forward into later periods through the purchase memory mechanism. Together these forces create inertia in the market that make it difficult to force switching of customers between brands. The improvement in predictive accuracy compared to the model without interactions is limited, but potentially of practical importance — the inclusion of interaction network appears to be of some use in explaining the choices of the individual agents and points to recommendations or social pressure as being a relevant factor. This suggests that managers in this field could make use of marketing mechanisms which accelerate the process of social pressure — for example by using shareable coupons or social media as part of their promotional activity.

- *Broader methodological issues* — From a methodological viewpoint, the chapter illuminated a number of issues relating to modelling social pressure using an agent based approach. As discussed in Chapter 2, there is an implicit assumption in equation based models that social agents interact not with each other, but with abstract economic objects like price vectors and unemployment rates [12]. In the absence of a known set of links between agents, the social network structure used in this application is essentially an abstract object, and a more specific set of links may create the basis for testing sharing hypothesis which further improve the prediction accuracy of the model. The role of social pressure in the simulation differs to that in Chapter 6 in several respects. Whereas the data collected in Chapter 6 was collected at an aggregate, *brand post* level, and carried no individual or time series information, in this application behavioural data was available for individuals over time. Since the choice model is estimated separately for each individual this allows for heterogeneous responses to social pressure across the individuals. In addition the fact that data is available over time means that social pressure can be simulated across time periods, rather than just within a single iteration as in Chapter 6. This reduces the influence of the issue of order dependence discussed in Section 6.8, since the social pressure effect is not solely determined in a single iteration.

  As reviewed in Section 2.4.4.4 an equilibrium value is a steady state that remains constant over time, although a fundamental change may occur which moves the process to a new equilibrium. It may be possible that given the stabilising effects of the agent's memory and purchase cycle mechanisms, sustained stimulus over a number of iterations is required to create the kind of level shift Huckfeldt [97] refers to. This may be the case in markets where user's preferences are less established, or there are fewer specialised niche products appealing to individual tastes.

  As discussed in Section 2.4.4, a number of existing studies have used learning methods to learn models of individual agent behaviour. Existing individual based statistical and data mining modelling approaches that are already being successfully applied to individual level observed data, for example in the microsimulation of economic behaviour [187] and credit scoring [87] often ignore network effects, with social behaviour modelled at the level of the individual, with each person considered to be independent and predictions about their behaviour made separately — even where data exists for the behaviour of multiple people in the same environment [208]. This chapter builds on these existing methods by incorporating elements of social interaction.

- *Applied marketing research* — From an applied marketing research perspective, the framework offers a platform for understanding why a company's sales converge towards an equilibrium value in a competitive environment, particularly where the constituent entities in the market are making decisions based on product features and habit. In addition, the ability to analyse the heterogeneous tastes of the individual agents in this application provides a way of understanding how preferences are related across the population, and also provides a basis to uncover sub-groups of agents with similar preferences. This information could be used to help design new products for particular market niches.

The next chapter concludes this thesis by reviewing and summarising the research conducted, considering its implications, and identifying some areas of work that could develop this research in the future.

Chapter 8

Summary and Conclusions

8.1 Introduction

This chapter concludes the thesis by summarising the research, evaluating its implications and suggesting additional work that could be carried out to extend and develop this research in the future. In Section 8.2 I summarise the research and relate it back to the objectives and contributions to the state of the art that were laid out in Chapter 1. Then in Section 8.3 the results of the research are evaluated, taking into account the type and quality of the empirical data that was used, and other possible methodologies that could have been deployed. Following that, in Section 8.4 I discuss some ways of extending this research in the future.

8.2 Contribution

As stated in Chapter 1, this thesis aims to solve the problem of selecting an agent based model specification and the theory that it implies from the many candidate specifications that may exist. The approach proposed was to develop an automated theory and model selection system. In Chapter 3 I created a simulation to compare the utility of automating EBM and ABM modelling approaches under different conditions of social interaction. I found that a common, simple automated approach works well for both EB and AB modelling methods at low levels of social interaction, suggesting that although the forms of these two types of model are different there are common keys to success in both.

I then proposed, in Chapters 4 and 5, methods in five key areas which I suggest are necessary components for a fully functional automated theory selection system using ABM:

1. **Theory representation** — I propose an ontological theory representation which is conformable with the specific characteristics of an agent based model, including emergent features. This builds on work on theory representation frameworks that have been proposed specifically for ABM, including MIMOSA [152] and a number based on the INGENIAS framework [175, 69]. My contribution is to extend these to include candidate relationships between the factors and the agents, multiple possible candidate elements for each theory element, and to integrate the ontology into a theory induction framework rather than a model development platform.

2. **Mapping theories to models** — I developed a flexible rule based system that allows the specification for a theory to be translated into a representation in a model.

3. **Model and theory scoring and criteria** — I propose a method for scoring the elements of the theory according to different criteria and weighting the set of inputs into an overall score.

4. **Model search mechanism** — I consider various methods for searching the theory and parameter space and propose a genetic algorithm as the most effective search procedure. From the calibration standpoint, this approach builds on work done by Fabretti and others in calibrating agent based models [61, 60]. In particular, I advance this work by designing a customised representation of a genetic algorithm which is better suited to agent based models through use of a non-binary alphabet.

5. **Interpreting a model in terms of a theory** — I developed a simple rule based theory interpretation system that can provide the basis for an autonomous action selection process.

Combining all of these elements creates a theory induction system which builds on work by Clark and others in equation based modelling contexts [40], and I claim advances the state of the art in structured agent based model testing and agent based theory induction. In addition, the approach applies the concepts put forward by Delre *et al.* [47] and North *et al.* [154], discussed in Section 2.6.4.2 regarding modelling of social diffusion. It also builds on the practical applications of Cha et al. [35], Sun et al. [190] and Yu and Fei [210] by applying an aggregate level model to an online social diffusion process and linking the drivers of diffusion to specific elements of online content.

By introducing a formal, replicable theory representation, this approach provides a high level of standardisation. Through its integrated search process, it allows for a consistent approach to validation. In addition it widens the range of modelling methods that can be accommodated under the agent framework specified by the theory representation. It also constrains the level of feedback and emergence that can influence the model's dynamics through the imposition of a maximum bound on the social interaction coefficient.

8.3 Evaluating the Results

The approach outlined proposes a system for learning theories from data, and each instance of the approach involves one main input — data — and two main outputs that could be evaluated — one a theory and the other a model, each with its associated characteristics. In the sections below I will consider a range of methods and criteria that could be used to evaluate the results in terms of these three elements. I will also consider ways of assessing the practical utility of the work overall.

8.3.1 Data

The theories and models that used in this research take as inputs a set of empirically observed historic data, recorded over time at either individual or aggregate level, which I take as a representation of the real world outcomes from the real-world-data-generating process.

The data used in this research came from two main sources:

- The data that is used in Chapter 6 — represents electronic records of the online actions of individuals and corporations relating to the allocation and display of online content to users. The data represents the actions of all of the individuals involved, rather than a sample. This data is described further in Section 6.6.
- In Chapter 7, the data used is collected from a sample of individuals, with each individual keeping records of their own purchasing behaviour by scanning the details of the purchased items into an electronic system. There is more information about this data in Section 7.5.

The validity of the theories and models derived depend on the assumption that this data is an accurate representation of this process. In the sections below I evaluate the data that used in terms of its ability to reflect the target empirical process in terms of the data collection method and possible sampling and recording biases. I then review some alternative methods that could have been used to collect this type of data for validation purposes.

8.3.1.1 Data Collection Bias

Data collection for any research purpose is susceptible to possible biases which may limit the degree to which they reflect the target data generating process, and therefore disrupt their suitability for the research objective. These biases are partly related to sampling — from which subset of the total population the data is collected, and the data collection instrument. Figure 8.1 illustrates these two main issues.

8.3.1.2 Sample Selection Bias

The data that used in Chapter 7 is based on a sample of the total purchasing population. If the data collected from the sample is intended to represent the characteristics of the target population, Fowler suggests that it is a fundamental premise of the survey process that by describing the sample of people who actually respond, one can describe the target population [64]. If the sample does not reflect the characteristics of the overall population, any conclusions drawn from the sample cannot be considered to hold for the population overall. In deriving a theory based on the sampled data used in Chapter 7 assumption is made that the data is representative of the total population, although if the sample is not representative, it may in reality be a theory of the sampled group.

8.3.1.3 The Data Collection Instrument

Regardless of whether or not the data is sampled, the data collection process can only be considered to be collecting the answers to the intended questions if the responses accurately reflect the characteristics that the researcher intends to measure. Fowler [64] suggests that it is a fundamental premise of the research process that the answers people give can be used to accurately describe their characteristics. Podsakoff[164] identifies a number of biases in this kind of research, including:

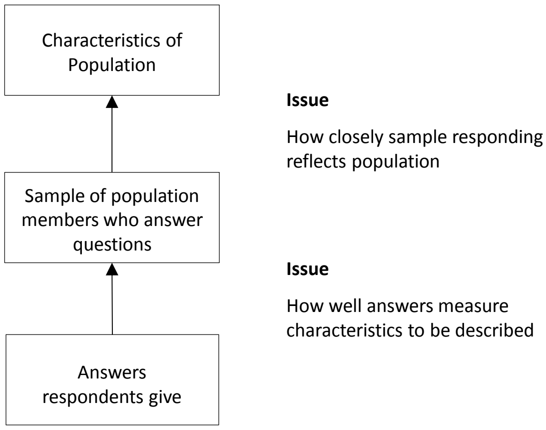

Figure 8.1: Sources of survey error — from [64]

- **Acquiescence biases** — this refers to the likelihood for respondents to agree or disagree with questionnaire items independent of their content.
- **Item social desirability** — this refers to the possibility that some items may be presented in such a way as to reflect more socially desirable attitudes, behaviours, or perceptions.
- **Measurement context effects** — like the framing effects discussed in Section 2.6.1, these refer to any variation in the research outcome that is produced from the context in which the measures are obtained.

The data used appears to be robust to these specific problems because the data collection methods used do not require a qualitative response. However, it is possible that the data is subject to other biases, for example that users of a social network may be reluctant to indicate their actual feelings about a post because their response is observed by their peers, or that an individual may not scan a particular purchase because they feel it reflects poorly on them.

8.3.1.4 Alternative Data Collection Methods

There are other ways of collecting data about such events, which could have been deployed to validate the accuracy of the data used. For example Hague *et al.* [86] review a number of qualitative research methods that are typically employed in marketing to elicit information from consumers. For example focus groups involve a meeting of a small number of selected people who have some experience of the topic with discussion guided by the moderator — these groups may provide useful insight but yield qualitative rather than quantitative information. Depth interviewing provides a way of exploring the thoughts of a single individual by using a loosely structured interview with open-ended questions, while questionnaires usually consist of a combination of behavioural and attitudinal questions [86].

These data collection methods can be executed through a number of means. Hague [86] suggests that face to face interviews are good for building rapport, observing respondent reactions, and for using body language. Face to face interviewers can also use visual stimuli [86]. Telephone interviews are faster and typically incur lower research costs than face to face interviews, but are not as useful if the subject of the research is something that needs to be shown to the subject. Online surveys differ from telephone interviews in the sense that they are self-administered. Grandcolas [83] argues that online surveys can be more appropriate than telephone interviews when the research topic requires visual rather than aural stimuli. Ethnography involves applying multiple data collection methods, ranging from surveys to observational data, video tapes, photographs and recordings. Ethnography involves the researcher having prolonged direct contact with members of the target group in an effort to look for rounded, holistic explanations [82].

The methods described above represent alternative means through which information about the real-world data generating process could have been collected. In addition to validating the accuracy of the data used, some of these methods could have been used to obtain more qualitative information about the context of the events, for example by asking the individuals involved about the reasons behind their choices.

8.3.2 Model

The models were evaluated in a number of ways:

- **Quantitative evaluation of fit to historic data** — One way of looking at the validity of a model is by considering its fit to historic data, although there may also be cases where models can be useful without this, serving instead as what Epstein calls illuminating abstractions [58]. The fit of the models I derived was evaluated using the R-Squared measure, finding that the fits of the models on historic data were 66% for the model obtained in Chapter 7 and 63% for that obtained in Chapter 7. In Chapter 2, several criteria were reviewed that could be used to assess the quality of a model, including the BIC and Adjusted R-squared. These measures quantify the degree of fit achieved against the historic data but also penalise less parsimonious models by taking into account the number of parameters required to achieve it. These and other measures of fit could also have been used to evaluate the models.

- **Comparison with other types of model** — another way of validating model output is to compare the output with other modelling approaches. I compared the model I obtained in Chapter 6 with the results of a linear regression model using the same data, which had an R-Squared of 56%. In Section 2.3.2 various alternative types of model were considered that could be used to validate the outputs of the agent based approach, including systems dynamics, cellular automata and equation based models.

Since equation based models are currently the most widely used type of model in this field [193], a comparison with linear regression was appropriate, but the models could also have been compared with other approaches.

- **Quantitative evaluation of predictive accuracy** — a further way of assessing a model's quality is by evaluating its predictive power. Hassan [90] suggests that forecast accuracy can be considered as one of the guiding criteria for model construction, and Thompson and Derr [197] argue that scientists almost universally agree that the predictions are what demonstrates a model's heuristic power. Epstein on the other hand argues that prediction is not a necessary property of a good model, using the example that evolution may be broadly accepted as an explanation of speciation, but that it is impossible to predict next year's flu strain.

  A criterion that is often used for gauging this predictive power is degree of forecasting accuracy. Forecasting accuracy can be measured in a variety of ways, some quantitative and others qualitative. From the point of view of creating objective measurements, quantitative measures based on the forecast error are often used. The forecast error is the difference between the actual value of the target variable in a particular time period and the predicted value, for example:

  $$e_t = Y_t - Yhat_t \tag{8.1}$$

  where e_t represents the forecast error, Y_t represents the actual value of the variable and $Yhat_t$ represents the predicted value.

  There are many different ways to aggregate the error in t over time [90]. Each resulting error measure has different properties. The forecasting accuracy of the model developed in Chapter 7 was tested and it was found that the classification accuracy of the model in the holdout period was 53%. This suggests that there was some level of over-fitting in the calibration period. However, forecasting accuracy may not necessarily be a consistent measure — since a model may have good predictive power in principal but poor forecasting accuracy if the initial conditions are inaccurately specified, or if predictions for the values of the exogenous variables are themselves inaccurate.

8.3.3 Theory

As discussed in Section 2.5.2, there are a variety of possible criteria for evaluating theories beyond empirical considerations, including consistency with itself and other accepted theories, simplicity, and the ability to reveal new research findings [119]. Ultimately the exact criteria used to assess a particular theory may be context-dependent, but some of the criteria by which a theory can be evaluated include:

- **Model quality** — As discussed above, the models obtained in Chapters 6 and 7 exhibited a reasonable level of fit to empirical data. Thompson and Derr argue that a good model implies the existence of theoretically significant properties or behaviours of the modelled phenomenon [197], the validity of the theory is therefore implied by the quality of the model. However Goldfarb et al. [79] consider theories to be empirically equivalent when they have identical empirical consequences, and suggest that since a trivial change to a theory will create a new theory but the empirical consequences will remain the same, there are an indefinite number of rival theories that may be created. This suggests that a theory should not be assessed purely in terms of its empirical consequences.

- **Compliance with existing theory** – In Section 2.5.2 a number of criteria for evaluating models were discussed in terms of their conformity with existing theory. Because the theory hypothesis space used is constrained, some degree of compliance with existing theory is assured. But in this kind of theory based specification the acceptability of the model in its domain depends on the credibility of the theory being applied — if the theory is later discredited then the model and its outputs are also subject to doubt. The overall structure allowed for and provided heterogeneous results across individuals. In Chapter 6 heterogeneity was incorporated in the distributions of preferences and behaviours across individuals, whereas in Chapter 7 the model were calibrated individually for the different agents.

- **Explanatory power** — I argue that the approach proposed offers more explanatory power than an equation based alternative, in the sense that it allows the consideration of a distinct influence of external forces on sharing and buying behaviour separately. Also, more complex interaction structures and transmission methods were tested than could have been represented in an EBM. The theories derived from the data were generative, in the sense that the data generating mechanism was specified at a micro-level and made up of individual level behaviour. As discussed in Section 2.4.4.5, there has been much debate in the literature about what constitutes an explanatory model, and how explanation relates to causality [85, 80, 41]. One way of validating the potential of a model for causal explanation is to design a real-world experiment that can test its implications. Goldthorpe [80] suggests that this is a form of consequential manipulation and is commonly used in practical sciences like medicine and agriculture. The causes are things that can be introduced as treatments in experiments — X is manipulated with all other variables controlled for and the outcome Y is systematically observed.

- **Comparison with theories derived from other methods** — As well as primarily quantitative approaches to theory development, such as the one used, there are also alternative, qualitative methods. Theories could be developed in parallel using qualitative methods, and these theories could be compared with the theory derived. For example Goulding [82] describes the main elements of the grounded theory approach to theory development using qualitative research, in which the researcher begins the study without exhausting the existing literature. She suggests that the theory the researcher develops as the study progresses should lead them to existing theories and literature that are relevant to the emerging concepts they see in the data. They should analyse interview texts, noting provisional themes, and compare these with other texts to ensure consistency and identify cases that are contradictory. The researcher then attempts to identify concepts that may offer an explanation of target phenomenon under study. The theory is then written up and integrated with existing theories to test its relevance and describe any new developments that are implied [82]. However Kaplan and Maxwell [108] argue that in qualitative data analysis there is a strong pressure to ignore data that does not fit existing theories, and that equal attention should be paid to data that supports an existing theory and data that refutes it. They suggest that in these cases the contradictory evidence should be presented and readers allowed to reach their own conclusions.

8.3.4 Practical Utility

Any theory, regardless of how it was derived, could be evaluated in terms of its practical utility. I claim that the results provided by the system proposed are useful in the sense that they provide the kind of insight that could be used to guide marketing policy decisions. This claim is based on comparison of my results with the kind of results obtained in the existing literature, as reviewed in Chapter 2 — for example the results provide information on the relative importance of the instruments reviewed in Section 2.6.1. The practical utility of the approach proposed could be further assessed by doing research amongst potential end users, such as marketing managers and planners. One way of doing this would be to allow end users to trial the system and then use some of the techniques discussed in Section 8.3.1.4 to compile information about their experiences. Li *et al.* [126] describe an evaluation of their WebDigital digital marketing system. They selected respondents based on their experience of using on-line systems and their familiarity with making digital marketing decisions. The participants used WebDigital to develop and formulate digital marketing strategies for their own cases within global markets through digital channels. After using WebDigital, they were then asked to complete an evaluation questionnaire with closed and open-ended questions. Kaplan and Duchon [107] conducted research into the utility of a new laboratory information software system by conducting fieldwork consisting of open-ended interviews, observation, and data analysis of survey questionnaire responses. They also used quantitative methods to collect and analyse data from survey questionnaires.

There was insufficient time and resource to carry out a survey to validate the results of this research in this way.

8.4 Future Work

The research that I have completed suggests a number of further areas of future work that could include research on:

- **Incorporating other emergent features** — As discussed in Chapter 2, Bedau [20] identifies weak emergence as an outcome that is derivable from the underlying micro component's facts, but only by simulation. In the theory specification and associated models that were developed, I looked at weakly emergent aggregate events that are dependent on the interactions of individual entities. Including other dynamics into the framework may also be interesting, and some of these dynamics could improve the level of fit of ABMs against empirical marketing data — in particular emergent aggregates that appear over time as a result of temporal rules and characteristics at agent level. An example of this could be to look at the consumer stockpiling behaviour in response to price promotions – in which consumers increase their inventories either by buying a higher quantity than normal or by buying earlier than they would have otherwise done. Other studies [153] have shown that purchase acceleration can cause pronounced changes in market share, consisting of a positive short-term effect, followed by the a longer-term, negative effect on sales. Purchase deceleration, in which consumers deplete their inventory below normal levels in anticipation of an upcoming promotion, could also be built into individual agent's behaviour.

- **Degrees of freedom** — It was sometimes difficult to calculate the real degrees of freedom in the agent based models that I was working with. Since increasing degrees of freedom typically leads to an increase in model fit, when comparing fit with other types of model, a method for calculating the real degrees of freedom in an ABM would be advantageous. Whilst in EBM the number of degrees of freedom typically relates directly to the number of parameters estimated, it is sometimes difficult to know how many parameters are active in an ABM. This is especially the case when the parameters are nested in complex dynamics or behaviours — for example when rule based behaviour or cognitive architectures are used.

- **Validation at different levels of aggregation** — In Section 2.4.4.2 I reviewed the debate about the level of aggregation at which agent based models should be validated. As part of my research into calibrating agent based models at individual and aggregate level, issues arose relating to potential conflict between success criteria at these different levels. For example, if the objective function of the calibration process is to fit the aggregate data series by calibrating agent parameters individually, an issue of order-dependence arises since the agents who are later in the queue for calibration are more likely to be parameterised so that the predicted outcome at individual level helps the aggregate fit, at the expense of micro-level fit. More work could be done on solving and understanding this aggregation problem and how the conflict between achieving fit at micro and macro level can be solved.

- **Incorporating new concepts** — The theory development approach that I use is constrained in the sense that any theory that is learnt, is learnt within the framework of an existing set of concepts. Unlike the grounded theory approach described in Section 8.3.3, new concepts cannot be added to theory that have not already been conceived before the research process is initiated. There are risks to defining a constrained hypothesis space — as Luger points out, the commitments made within a learning scheme determine to a large extent the results one can expect from it [135]. In the learning approach that I propose, it is the responsibility of the domain specialist to ensure that the constrained hypothesis space is large enough to contain the solution to the problem being learnt, since by eliminating some candidates it becomes impossible to learn any of the concepts that relate to those hypotheses. If the process is unable to find a satisfactory solution within the constraints and data specified then the user will need to consider revising their theory, including using different or additional exogenous data and different sets of constraints. Kuhn argues that scientific discovery often involves a paradigm change the genesis of a new theory. He suggests that for as long as possible people will resist evidence that is unexpected and conflicts with their past experience [88]. A fuller approach would support the ability to learn new concepts and relationships to reflect the possibility of this kind of paradigm change.

- **Investigating dynamics of equilibrium attainment** — In Section 2.4.4.4 I reviewed existing research into the time-series properties of data produced by agent based models and in Chapter 7 I found that under certain simulated conditions the level of sales for each of a number of brands was robust to reasonable interventions. The framework that I developed as part of the research could be used to develop this by studying how a company's sales converge towards an equilibrium value — a steady state that remains constant over time — in a competitive environment. Epstein [58] argues that there are at least three scenarios in which an equilibrium will not be achieved — the phenomenon is a non-equilibrium dynamic, the time-scales over which equilibrium might be reached are unrealistic, equilibrium exists but is unattainable. In order to examine these possibilities in a market environment, the multi-player simulation that I developed in Chapter 3 could be extended to reflect a more complex decision structure for each participant, and with the interactions between the players extending over greater number of iterations.

- **Parameter significance** The system that I have proposed currently outputs a measure of the overall fit of the model. In the context of theory derivation, it would be useful to also have a measure of the significance of the individual parameters contained in the model, measured at some benchmark significance level — e.g. 5% [110]. There has been some research focussing on statistical validation of agent based models, including the theory of maximum likelihood approaches in agent based simulation [115] and its applications in computational finance [136, 4, 127], macroeconomics [39], geography [103], and diffusion analysis [67]. These approaches could be used to extend the system to use a maximum likelihood framework to test the significance of the different elements in the final theory.

- **Developing the theory representation** — The ontological theory specification that I propose contains elements at a certain level of abstraction. Zuniga considers what makes any one thing an economic object [211]. She suggests that individuals view the world as consisting of economic facts, but there are more fundamental facts, and that economic objects are a product of both subjective and objective factors — beliefs about them and their intrinsic properties. The ontology that I define incorporates beliefs, in the sense that the agent's perceptions and preferences are represented, but the ontological structure could be developed to include the possibility of changing utility for products over time. In addition, some of the high-level concepts, such as memory, could be developed to incorporate more fundamental elements, such as long and short term components.

- **Variance decomposition** — The outcome of the system that I have developed contains two principal types of variation, variation that is exogenous — directly caused by external factors, and that which is endogenous — that is primarily caused by dynamics within the model. In this sense the outcome is similar to studies such as the Artificial Anasazi model [58, 102], which aims to manipulate parameters to maximise fit against an expected outcome at an aggregate level, and contains a combination of internal and external dynamics. However Janssen [102] argues that in the case of the Artificial Anasazi model, the internal dynamics of the model make little contribution to the fit, and that the model acted largely as a smoothing function of the exogenous data, with limited contribution from endogenous dynamics. A facility for understanding this variance partition — which parts of the variance are due to which effects, could be added to the modelling system, perhaps by testing for first order and global impact on the outcome for each of the theory elements [174, 173, 129].

- **Application to other domains** — The method could be adapted and tested in different domains — especially those that are more likely to exhibit less defined or predictable emergent behaviour. Some applications to other areas could include parameterising the characters in video games, based on real world behaviour. Gillies *et al.* [77] use a similar approach in calibrating the behavioural parameters of a game AI character to reflect the responses and movements of a real actor to stimuli in the environment. The data they used was collected from video capture of the actor's movements and interpreted using principal components and clustering methods. The interpreted movements were combined with a corresponding dataset of events in the environment that the actor was considered to be responding to. The system could be modified so that the theory representation reflects concepts and rules related to the possible behaviours of the associated character in the game, and their interactions.

8.5 Conclusion

This chapter concludes the thesis by reviewing and summarising the research conducted in Section 8.2, considering its implications in Section 8.3, and identifying some areas of work that could develop this research in the future in Section 8.4.

Chapter 9

Appendix1

This appendix shows some statistical summaries for the results from the genetic algorithm implementation in Chapter 6. The search procedure was repeated 30 times to illustrate its stability across multiple runs. There are five exhibits in this section. Figure 9.1 illustrates the progress of the score variable, for a randomly selected iteration, over 100 generations. The score proceeds towards a minimum, initially in large steps, then smaller moves until it stabilises on the 70th generation and remains at a constant level. Figure 9.2 shows the distribution of the score variable, for all of the chromosomes evaluated. In other words it is a summary of the evaluated score variable for each of the models evaluated across all populations and generations for the 30 iterations completed. Figure 9.3 shows the distribution of the best scores for each of the 30 iterations. There is a relatively small range of scores, indicating that the search procedure arrives at a very similar score value on each iteration. The final figure, Figure 9.4, shows the correlation between the parameter estimates across the 30 iterations.

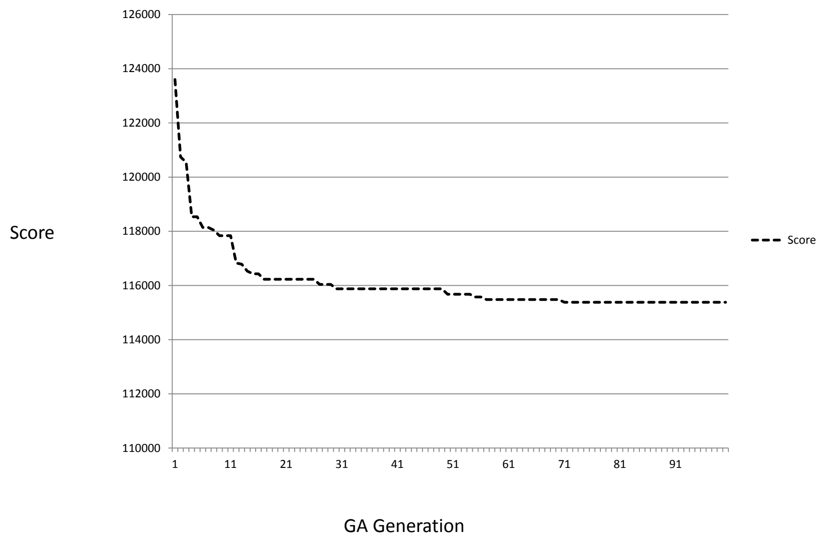

Figure 9.1: The development of the calculated score over 100 generations of the genetic algorithm

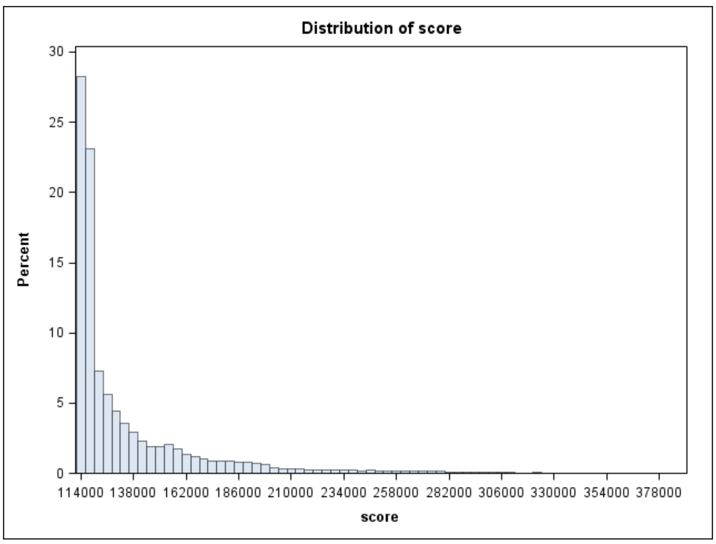

Figure 9.2: The distribution of the score variable across all populations and generations for all of the 30 iterations

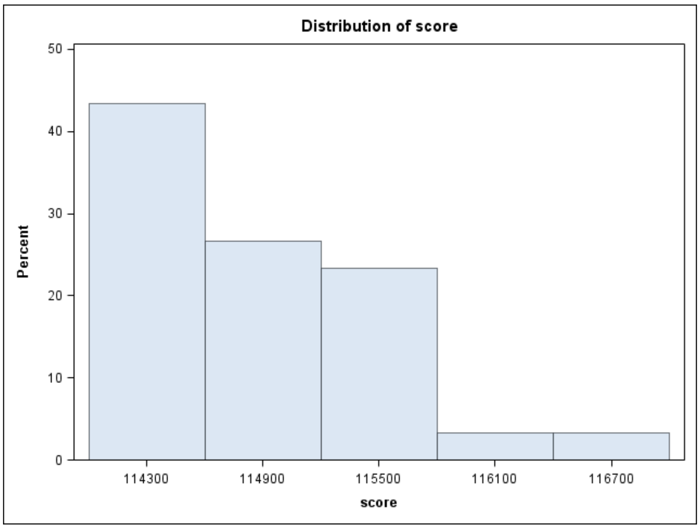

Figure 9.3: The distribution of the best score variable across all of the 30 iterations

	p_Intercept	p_coffee	p_days_since_last	p_exmark	p_frappucino	p_hour	p_latte	p_qmark	p_textlength	p_social_pressure	p_weekday_1	p_weekday_2	p_weekday_4	p_weekday_5	p_weekday_6
p_Intercept	1.00000	0.03015	0.54490	0.04303	0.14512	-0.25449	0.09063	-0.24067	0.04227	-0.02321	-0.37941	0.22313	0.10179	-0.04865	0.10991
p_coffee	0.03015	1.00000	-0.14292	-0.20843	-0.04227	-0.08314	-0.11115	-0.60680	0.04773	0.11614	-0.14528	-0.23956	0.18633	0.27828	-0.01519
p_days_since_last	0.54490	-0.14292	1.00000	-0.04690	-0.04489	-0.19037	0.26142	-0.01804	0.19468	-0.18337	-0.41148	0.17780	-0.01770	-0.59863	-0.00349
p_exmark	0.04303	-0.20843	-0.04690	1.00000	0.16394	-0.14070	-0.13137	0.03883	0.08094	-0.00771	0.20945	-0.14793	0.11408	-0.06574	-0.25592
p_frappucino	0.14512	-0.04227	-0.04489	0.16394	1.00000	0.02596	-0.10995	0.01116	-0.00616	-0.01426	0.01938	-0.08365	0.11107	-0.01923	-0.01468
p_hour	-0.25449	-0.08314	-0.19037	-0.14070	0.02596	1.00000	0.07365	-0.11639	-0.20495	-0.84266	0.48125	0.04179	-0.12712	-0.09800	0.10834
p_latte	0.09063	-0.11115	0.26142	-0.13137	-0.10995	0.07365	1.00000	-0.00517	0.16997	-0.20801	0.06965	0.45013	-0.40071	0.03982	0.19840
p_qmark	-0.24067	-0.60680	-0.01804	0.03883	0.01116	-0.11639	-0.00517	1.00000	0.08058	0.17161	0.23280	-0.11138	-0.36086	-0.12682	-0.25479
p_textlength	0.04227	0.04773	0.19468	0.08094	-0.00616	-0.20495	0.16997	0.08058	1.00000	-0.15962	0.04581	0.38817	-0.74898	-0.23980	-0.15217
p_social; pressure	-0.02321	0.11614	-0.18337	-0.00771	-0.01426	-0.84266	-0.20801	0.17161	-0.15962	1.00000	-0.44696	-0.30509	0.27951	0.23493	-0.05640
p_weekday_1	-0.37941	-0.14528	-0.41148	0.20945	0.01938	0.48125	0.06965	0.23280	0.04581	-0.44696	1.00000	0.08978	-0.19213	0.30315	-0.00206
p_weekday_2	0.22313	-0.23956	0.17780	-0.14793	-0.08365	0.04179	0.45013	-0.11138	0.38817	-0.30509	0.08978	1.00000	-0.34346	-0.00137	0.15876
p_weekday_4	0.10179	0.18633	-0.01770	0.11408	0.11107	-0.12712	-0.40071	-0.36086	-0.74898	0.27951	-0.19213	-0.34346	1.00000	0.18379	0.06510
p_weekday_5	-0.04865	0.27828	-0.59863	-0.06574	-0.01923	-0.09800	0.03982	-0.12682	-0.23980	0.23493	0.30315	-0.00137	0.18379	1.00000	0.07241
p_weekday_6	0.10991	-0.01519	-0.00349	-0.25592	-0.01468	0.10834	0.19840	-0.25479	-0.15217	-0.05640	-0.00206	0.15876	0.06510	0.07241	1.00000

Figure 9.4: The correlation between estimated parameters across all of the 30 iterations

Chapter 10

Appendix2

The code in this section relates to the implementations in Chapter 3.

SAS Code

```
 data history;

%macro master_manage;
%do p= 1 %to 100;
X "cd C:\Users\rstratton\workspace\test";
X 'javaw -cp "C:\Program Files (x86)\NetLogo 5.0.3\netlogo.jar";. master_nlogo';
run;


PROC IMPORT OUT= WORK.in_
 data
          DATAFILE= "C:\Users\rstratton\Dropbox\coffee project\agent model\plot.csv"
          DBMS=CSV REPLACE;
    GETNAMES=no;
    DATAROW=1;
RUN;


 data in_data;
retain period 0;
set in_data;
                    period=period + 1;
                    iteration=&p;
                    man1_sales= var1;
                    man2_sales= var5;
                    man3_sales= var9;


 data history;
set history in_data;
run;


PROC EXPORT DATA= WORK.history
          OUTFILE= "C:\Users\rstratton\Dropbox\coffee project\agent model\history_archive.txt"
          DBMS=TAB REPLACE;
    PUTNAMES=NO;
RUN;
%end;

proc sort data=history;
by iteration;

proc means data=history;
by iteration;
run;

%mend;

run;

%master_manage;
run;

%macro run_model;

X "cd C:\Users\rstratton\workspace\test";
X 'java -cp "C:\Program Files (x86)\NetLogo 5.0.3\netlogo.jar";. run_nlogo';
run;
%mend;
%run_model;

options  nosymbolgen nomacrogen nonotes nomprint;
%let tuning_phase=0;
%let tuning_threshold=10;
%let tuning_period=0;
%let tuning_length=10;
%let groupcount=3;
libname d 'C:\Users\rstratton\Dropbox\coffee project\agent model';
%let file_groups= C:\Users\rstratton\Dropbox\coffee project\agent model\ga_input_parm_groups.csv;
options  noquotelenmax;
%include "C:\Users\rstratton\Dropbox\coffee project\agent model\ga_code_import_parm_groups.sas";
%include "C:\Users\rstratton\Dropbox\coffee project\agent model\ga_code_Run_Optimisation teeth.sas";
%macro import_target_variable;
%mend;

 data score;

 data d.best;
%macro looper;
*%import_groups;
%import_target_variable;
%best_guess;

 data d.targetVariable;
infile 'C:\Users\rstratton\Dropbox\coffee project\agent model\sas_target_variable.csv' delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=1 ;
input y1 ;
run;

 data champion_model;
set champion_model ;
/*
```

```
 data champion_model;
set champion_model d.bestyet;
zweight=250;
*/
%do vk= 1 %to 2;
%put &loop;

 data _null_;
retain rowcount 0;
set champion_model;
length varb $10000;
rowcount=rowcount+1;
counter=0;
varb="";
%do i=1 %to &groupcount;
if &&groupname&i="" then delete;
if &&groupname&i=. then delete;
%end;
			%do i=1 %to &groupcount;;
						call symput('myvar_'||left(compress(&i))||'_'||left(trim(rowcount)),&&groupname&i);
						call symput('model_count',left(trim(rowcount)));
			%end;
run;

 data score;

%do i= 1 %to &model_count;
			%run_sim;
			%store_results;
%put &i;
%end;
%cross_over;
%evaluate_history;

 data top;
set score(obs=1);
call symput('score',score);
run;
%put &score;

 data d.best;
set d.best top;
run;
%end;

 data d.best;
set d.best;
where score ^=.;

proc sort data=d.best;
by score;

 data final;
set d.best(obs=1);
run;

PROC EXPORT DATA= WORK.final
         OUTFILE= "C:\Users\rstratton\Dropbox\coffee project\agent model\learning man2.txt"
         DBMS=TAB REPLACE;
    PUTNAMES=NO;
RUN;
%mend;
run;
%looper;
%Macro import_groups;


 data GROUPS ;
     infile "&file_groups" delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=2 ;
			informat VAR1 best32. ;
						informat VAR2 $50. ;
			format VAR1 best12. ;
			format VAR2 $50. ;
input VAR1 VAR2 $ ;
run;

proc sort data=groups;
by var2;
run;

 data groups_s;
set groups (where= (var2 ^= ' '));
length pregroup $255.;
retain n 0 g 0 pregroup 'ini';
if pregroup = var2 then
n=n + 1;
else do;
			n=1;
			g=g+1;
end;
call symput ("g"||trim(left(g))|| "v" ||left(n),trim(left(var1)));
pregroup = var2 ;
run;
```

```
%put &g1v30;

proc sql;
create table macrogroup as select var2 as var2, count(*) as count from groups_s group by var2 order by var2;

 data b;
set macrogroup;
retain counter 0;
if var2 ne '';
counter = counter + 1;
call symput ('groupname'||left(counter),var2);
call symput ('groupcount',left(trim(counter)));
call symput ('groupvarcount'||left(trim(counter)),left(trim(count)));

%mend;

%macro run_sim;

 data parms;
		%do z=1 %to &groupcount;;
				&&groupname&z ="&&myvar_&z._&i" ;
		%end;
run;

 data _null_;
 set  WORK.PARMS end=EFIEOD;
		file 'C:\Users\rstratton\Dropbox\coffee project\agent model\netlogo_est_parameters.txt' delimiter='09'x DSD DROPOVER lrecl=32767;
		do;
		%do z=1 %to &groupcount;;
				put &&groupname&z @ ;
		%end;
		end;
run;

 data mode;
mode=1;
X "cd C:\Users\rstratton\workspace\test";
X 'javaw -cp "C:\Program Files (x86)\NetLogo 5.0.3\netlogo.jar";. run_nlogo';
run;

 data WORK.abmOutput;
infile 'C:\Users\rstratton\Dropbox\coffee project\agent model\netlogo_out_history.csv' delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=1 ;
input y1_hat;
 run;
%mend;
run;
%macro store_results;

 data compare;
merge d.targetVariable abmOutput;
error1=(y1-y1_hat) * (y1-y1_hat);
error = error1 ;

proc sql;
create table sumOfSquaredErrors as select sum(error)as score from compare;
run;

 data record;
merge parms sumOfSquaredErrors;
run;

 data score;
set score record;
run;

%mend;

%macro import_target_variable;

 data d.targetVariable;
infile 'C:\Users\rstratton\Dropbox\coffee project\agent model\netlogo_out_history.csv' delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=1 ;
input y1 y2 y3;
 run;

%mend;
%macro best_guess;

 data champion_model;
length
%do i=1 %to &groupcount;
				&&groupname&i $200.
%end;
;
%do it=1 %to 32;
		%do i=1 %to &groupcount;
		%let random_no= %sysevalf(%sysfunc(ranuni(0))*&&groupvarcount&i,ceil);
				&&groupname&i = "&&g&i.v&random_no";

		%end;
		output;
%end;
run;
%mend;
%macro cross_over;
```

```
libname x "C:\Users\rstratton\Dropbox\coffee project\agent model\ga files";

%let split= %sysevalf(&groupcount/2,floor);
%let next= %sysevalf(&groupcount-&split);

 data score;
set score;
%do i=1 %to &groupcount;
        if &&groupname&i="" then delete;
        if &&groupname&i=. then delete;
%end;
run;

proc sort data=score;
by  score;

 data best;
set score(obs=1);

 data x.cross_over_candidates;
retain count 0;
set score(obs=10);
rand1=ranuni(0);
rand2=ranuni(0);
count=count + 1;
keep
rand1
rand2
count
%do i=1 %to &groupcount;
        &&groupname&i
%end;
;

 data x.mix1;
set x.cross_over_candidates;

proc sort data=x.mix1;
by rand1;

 data x.mix2;
set x.cross_over_candidates;

proc sort data=x.mix2;
by rand2;

 data x.mix1;
set x.mix1;
%do i=1 %to &groupcount;
        mix_1_&&groupname&i =&&groupname&i;
%end;
drop
%do i=1 %to &groupcount;
        &&groupname&i
%end;
;

 data x.mix2;
set x.mix2;
%do i=1 %to &groupcount;
        mix_2_&&groupname&i =&&groupname&i;
%end;
drop
%do i=1 %to &groupcount;
        &&groupname&i
%end;
;

 data x.cross_over_out1;
merge x.cross_over_candidates x.mix1 x.mix2;
%do i=1 %to &groupcount;
        myrand=ranuni(0);
        if myrand>.66 then &&groupname&i = mix_1_&&groupname&i;
        if myrand>.33 and myrand <=.66 then &&groupname&i = mix_2_&&groupname&i;
%end;

 data x.cross_over_candidates2;
retain count 0;
set score(obs=5);
rand1=ranuni(0);
rand2=ranuni(0);
count=count + 1;
keep
rand1
rand2
count
%do i=1 %to &groupcount;
        &&groupname&i
%end;
;

 data x.mix1;
set score(firstobs=10 obs=15);
rand1=ranuni(0);
keep
```

```
rand1
%do i=1 %to &groupcount;
            &&groupname&i
%end;
;

proc sort data=x.mix1;
by rand1;

 data x.mix2;
set score(firstobs=25 obs=30);
rand2=ranuni(0);
keep
rand2
%do i=1 %to &groupcount;
            &&groupname&i
%end;
;

proc sort data=x.mix2;
by rand2;

 data x.mix1;
set x.mix1;
%do i=1 %to &groupcount;
            mix_1_&&groupname&i =&&groupname&i;
%end;
drop
%do i=1 %to &groupcount;
            &&groupname&i
%end;
;

 data x.mix2;
set x.mix2;
%do i=1 %to &groupcount;
            mix_2_&&groupname&i =&&groupname&i;
%end;
drop
%do i=1 %to &groupcount;
            &&groupname&i
%end;
;

 data x.cross_over_out2;
merge x.cross_over_candidates2 x.mix1 x.mix2;
%do i=1 %to &groupcount;
            myrand=ranuni(0);
            if myrand>.5 then &&groupname&i = mix_1_&&groupname&i;
            if myrand< .5 then &&groupname&i = mix_2_&&groupname&i;
%end;

 data limit;
retain count 0;
set score (obs=10);
count=count+1;
call symput('rowcount',count);
threshold=ranuni(0);
run;

%do r=1 %to &rowcount;

            data x.mutation&r;
            set limit;
            if count=&r;
            %do i=1 %to &groupcount;
            myrand= %sysevalf(%sysfunc(ranuni(0)));
                        %let random_no= %sysevalf(%sysfunc(ranuni(0))*&&groupvarcount&i,ceil);
                        if myrand> threshold then &&groupname&i = "&&g&i.v&random_no";
                        %end;
%end;

 data x.mutation;
set
%do r=1 %to &rowcount;
            x.mutation&r
%end;
;

 data x.mutants;
length
%do i=1 %to &groupcount;
                        &&groupname&i $200.
%end;
;
%do it=1 %to 5;
            %do i=1 %to &groupcount;
            %let random_no= %sysevalf(%sysfunc(ranuni(0))*&&groupvarcount&i,ceil);
                        &&groupname&i = "&&g&i.v&random_no";

            %end;
            output;
%end;
run;
```

```
 data x.cross_overs;
set x.cross_over_out1 x.cross_over_out2;

proc sql;
create table x.cross_overs as select
                %do z=1 %to &groupcount;
                                &&groupname&z  ,
                %end;
count(*) as count, "0" as monkey from x.cross_overs
group by        %do z=1 %to &groupcount;
                                &&groupname&z  ,
                %end;
                monkey;

 data local;
set best;
do i=1 to 1;
%do i=1 %to &groupcount;
                adjuster=0;
                ran=ranuni(0);
                if ran >.7 then adjuster=1;
                if ran >.9 then adjuster=2;
                if ran <.3 then adjuster= -1;
                if ran <.1 then adjuster= -2;
                                                &&groupname&i = &&groupname&i + adjuster;
%end;
;
output;
end;
%if &tuning_phase=1  %then %do;

 data tp;
                tp=&tuning_period;
                call symput('tuning_period',(tp + 1));

 data local;
                set best;
                do i=1 to 10;
                %do i=1 %to &groupcount;
                if ranuni(0) > .5 then adjuster=0;
                ran=ranuni(0);
                if ranuni(0) > .9 then adjuster=ran;
                if ranuni(0) > .5 then adjuster=adjuster * -1;
                if ranuni(0) > .5 then adjuster=0;
                                                &&groupname&i = &&groupname&i + adjuster;
                %end;
                ;
                output;
                end;
%end;

 data champion_model_1;
set  best local x.cross_overs  x.mutants x.mutation;
run;

proc sql;
create table champion_model_2 as select
                %do z=1 %to &groupcount;
                                &&groupname&z,
                %end;
count(*), "1" as x
from champion_model_1
group by
%do z=1 %to &groupcount;
                                &&groupname&z,
                %end;
                x
;

 data champion_model;
set champion_model_2(obs=37);
%mend;
%macro evaluate_history;
%global static_history ;
%global tuning_phase;

 data evaluate;
set d.best;
                %do z=1 %to &groupcount;
                                lag_&&groupname&z =lag(&&groupname&z) ;
                %end;
static=0;
%do z=1 %to &groupcount;
                if lag_&&groupname&z = &&groupname&z then static=static + 1;
                static_score=static/&groupcount;
%end;

 data evaluate;
retain static_history 0;
set evaluate;
if static_score=1 then static_history=static_history +1;
if static_score<1 then static_history=0;
call symput('static_history',static_history);
run;
%if %sysevalf(&static_history>=&tuning_threshold) %then %do;
```

```
 data bb;
call symput('tuning_phase',1);
%end;
%mend;
run;

%put &tuning_phase;
```

Netlogo Code

```
breed[manufacturers manufacturer]
breed[consumers consumer]
extensions[shell]
globals [     man_1_imp1 man_1_imp2 man_1_imp3
  man_2_imp1 man_2_imp2 man_2_imp3

  man_1_media_one_grps
          man_1_coupons
          man_1_promotions

          man_2_media_one_grps
          man_2_coupons
          man_2_promotions

          man_3_media_one_grps
          man_3_coupons
          man_3_promotions

          links-list obs ob
          total_sec
          temp_sec
          report_sec
          intercept
          ad_coeff
          promo_coeff
          coupons_coeff

          man_no
          brand_1_sales
          brand_2_sales
          brand_3_sales

          brand_1_sales_sum
          brand_2_sales_sum
          brand_3_sales_sum
          ]
manufacturers-own[ad_belief promo_belief coupons_belief chosen_strategy man_id]
consumers-own [
          uses_per_week
          elapsed
          current_pack
          in_market?
          my_cycle
          brand_preference
          brand_buyer

          ever_active
          first_active

          node-id
          message?
          message_strength


          my_x
          my_y
          weight
          media_one_cons
          media_two_cons
          media_three_cons
          media_four_cons
          media_five_cons

          man_1_media_one_exp
                    man_2_media_one_exp
                              man_3_media_one_exp

          media_two_exp
          media_three_exp
          media_four_exp
          media_five_exp


          media_total_exp
          cum_media_total_exp
          remembered_exposures
          media_total_exp_man_1
          cum_media_total_exp_man_1
          remembered_exposures_man_1

          media_total_exp_man_2
          cum_media_total_exp_man_2
          remembered_exposures_man_2
```

```
          media_total_exp_man_3
          cum_media_total_exp_man_3
          remembered_exposures_man_3


          exposure

          exposure_man_1
          exposure_man_2
          exposure_man_3

          temp_memory

          temp_memory_man_1
          temp_memory_man_2
          temp_memory_man_3

          wom_exp

          cum_exp_count
          semiactive?
          active?

           active_man_1?
           active_man_2?
           active_man_3?

          brand_1_base_preference
          brand_2_base_preference
          brand_3_base_preference

          habit

          ad_1_effective
          ad_2_effective
          ad_3_effective

          utility_man_1
          utility_man_2
          utility_man_3

          last_purchase_man_1
          last_purchase_man_2
          last_purchase_man_3

last_brand_buyer
          ]

links-own [strength linkto linkfrom]
to setup
  __clear-all-and-reset-ticks

;MANUFACTURERS;
 let file "C:/Users/rstratton/Dropbox/coffee project/agent model/plot.csv"

   if file-exists? file
    [file-close
      file-delete file ]

  let file2 "C:/Users/rstratton/Dropbox/coffee project/agent model/grps.csv"

   if file-exists? file2
    [file-close
      file-delete file2 ]

     let file3 "C:/Users/rstratton/Dropbox/coffee project/agent model/sas_target_variable.csv"

   if file-exists? file3
    [file-close
      file-delete file3 ]

 ;CONSUMERS
ask consumers[ set cum_media_total_exp []]
  ask patches [set pcolor white]
  set-default-shape consumers "person"
open_file
import_parameters
import_agents
set man_no 0
repeat 3 [set man_no man_no + 1 create-manufacturers 1[set ad_belief 0 set promo_belief 0 set coupons_belief 0 set man_id man_no]]
import_links
  set man_1_media_one_grps 0
  set man_1_coupons 0
  set man_1_promotions 0

  set man_2_media_one_grps 0
  set man_2_coupons 0
  set man_2_promotions 0

  set man_3_media_one_grps 0
  set man_3_coupons 0
  set man_3_promotions 0
```

```
 set brand_1_sales 0
 set brand_2_sales 0
 set brand_3_sales 0


  ;import-drawing "logo3.bmp"

  ask consumers[

    set cum_media_total_exp []

    set remembered_exposures 0
]
end
to import_agents
   ;;Import the Agent Characteristics
  file-open "coordinates.txt"
  let incrementer 0
  let brand_incrementer 1
  while [not file-at-end?]
  [
    ;; this reads a single line into a three-item list
    let items read-from-string (word "[" file-read-line "]")
    ;show (list item 0 items item 1 items item 2 items)
    create-ordered-consumers 1
    [
    set my_x item 0 items
    set my_y item 1 items
    set node-id item 2 items
    set weight item 3 items
    set media_one_cons item 4 items
    set media_two_cons item 5 items
    set media_three_cons item 6 items
    set media_four_cons item 7 items
    set media_five_cons item 8 items
    set my_cycle item 9 items
    set brand_preference item 10 items
    set elapsed item 11 items
    set current_pack my_cycle

    set brand_incrementer brand_incrementer + 1
    if brand_incrementer > 3 [set brand_incrementer 1]
    set incrementer incrementer + .044
    if brand_incrementer = 1 [set brand_1_base_preference incrementer]
    if brand_incrementer = 2 [set brand_2_base_preference incrementer]
    if brand_incrementer = 3 [set brand_3_base_preference incrementer]
    set shape "person"
    set color (5 + (random 8)) * 10 + 4
    set size 1
    set label-color 2
    setxy my_x my_y
    ]
  ]
  file-close
end
to import_links

  clear-links

  if network_type = "None" [file-open "regular.txt"]
  if network_type = "Regular" [file-open "regular.txt"]
  if network_type = "Small World" [file-open "small_world.txt"]
  if network_type = "Random" [file-open "random.txt"]

  while [not file-at-end?]
  [let items read-from-string (word "[" file-read-line "]")
  ask turtle (item 0 items) [create-link-with turtle (item 1 items)  [ hide-link]]]
  file-close

  if influentials = "Yes"
  [file-open "influentials.txt"
   while [not file-at-end?]
   [let items read-from-string (word "[" file-read-line "]")
   ask turtle (item 0 items) [create-link-with turtle (item 1 items)  [hide-link]]]
   file-close
  ]


  if network_type != "None" [ask links [show-link set color 8]]
  if network_type = "Random" [repeat 5 [layout-spring consumers links 0.18 0.01 1.2 display]]
 ; if network_type = "Small World" or network_type = "Regular"  [repeat 100 [ask consumers [facexy my_x my_y fd .05 ]] display ]


end
to import_parameters
  file-open "parameters.txt"
  ;; Read in all the
 data in the file
  while [not file-at-end?]
  [
    ;; this reads a single line into a three-item list
    let items read-from-string (word "[" file-read-line "]")
    set ad_coeff item 0 items
    set coupons_coeff item 1 items
    set promo_coeff item 2 items
```

```
    ]
  file-close
;if ((item ticks media_one_grps) +  (item ticks media_two_grps) + (item ticks media_three_grps) + (item ticks media_four_grps) + (item ticks media_five_grps)) > 0
end
to open_file
let file "C:/Users/rstratton/Dropbox/coffee project/agent model/plot.csv"

    if file-exists? file
     [file-close
       file-delete file ]

end
to go

  if ticks > 51 [ stop ]

  ;MANUFACTURERS
   if ticks = 20 [learn_effects]

  ask manufacturers[select_strategy]
  deploy_policy
  monitor_sales



  ;CONSUMERS
;ask consumers [ad_assign_exposures]
  ask consumers [ad_evaluate]
 ; ask links [color-links]
;  ask consumers [deprecate]
  ask consumers [increment]
  ask consumers [set_pack_size]
  ask consumers [assess_utility]
  ask consumers [repeat_purchase]
  ask consumers [brand_switcher]
  update_aggregates
;  ask consumers [ recolor ]
  my-setup-plots
  sec
  do-cover-plots
write-to-file
  tick
  set ob ob + 1

end
to write-to-file

file-open "C:/Users/rstratton/Dropbox/coffee project/agent model/plot.csv"
;; assuming Windows machine
file-print (word(count consumers with [brand_buyer = "man_1"]) "," man_1_media_one_grps "," man_1_promotions "," man_1_coupons "," (count consumers with [brand_buyer =
"man_2"]) "," man_2_media_one_grps "," man_2_promotions "," man_2_coupons "," (count consumers with [brand_buyer = "man_3"]) "," man_3_media_one_grps ","
man_3_promotions "," man_3_coupons) ;; File is in writing mode
file-close
file-open "C:/Users/rstratton/Dropbox/coffee project/agent model/grps.csv"
;; assuming Windows machine
file-print (word( man_1_media_one_grps) " "  man_1_coupons " " man_1_promotions  " "  man_2_media_one_grps " " man_2_coupons " "  man_2_promotions " "
man_3_media_one_grps " " man_3_coupons " " man_3_promotions ) ;; File is in writing mode
file-close
file-open "C:/Users/rstratton/Dropbox/coffee project/agent model/sas_target_variable.csv"
;; assuming Windows machine
file-print (word((count consumers with [brand_buyer = "man_1"]))) ;; File is in writing mode
file-close
end
to increment

  ifelse elapsed >= current_pack
  [set elapsed  1 ]
  [set elapsed elapsed + 1 ]

end
to set_pack_size

  let ps 0
 ; if item ticks promotions > 0 [set ps item ticks promotions - 1]
  if elapsed = 1 [set current_pack my_cycle * (1 + ps)]

end
to assess_utility

  set ad_1_effective 0
  set ad_2_effective 0
  set ad_3_effective 0

  if active_man_1? = true [set ad_1_effective 1]
  if active_man_2? = true [set ad_2_effective 1]
  if active_man_3? = true [set ad_3_effective 1]

 ; set last_brand_buyer brand_buyer

    if last_brand_buyer = "man_1" [set last_purchase_man_1 last_purchase_man_1 + 50 ]
  if last_brand_buyer = "man_2" [set last_purchase_man_2 last_purchase_man_2 + 50]
  if last_brand_buyer = "man_3" [set last_purchase_man_3 last_purchase_man_3 + 50]
```

```
   set last_purchase_man_1 last_purchase_man_1 * habit_coeff
   set last_purchase_man_2 last_purchase_man_2 * habit_coeff
     set last_purchase_man_3 last_purchase_man_3 * habit_coeff



  set utility_man_1 40 + brand_1_base_preference + ad_coeff * ad_1_effective +   man_1_promotions * promo_coeff +   man_1_coupons * (-1 * coupons_coeff) + habit_coeff *
last_purchase_man_1 + sp * brand_1_sales
  set utility_man_2 40 + brand_2_base_preference + ad_coeff * ad_2_effective +   man_2_promotions * promo_coeff +   man_2_coupons * (-1 * coupons_coeff) + habit_coeff *
last_purchase_man_2 + sp * brand_2_sales
  set utility_man_3 40 + brand_3_base_preference + ad_coeff * ad_3_effective +   man_3_promotions * promo_coeff +   man_3_coupons * (-1 * coupons_coeff) + habit_coeff *
last_purchase_man_3 + sp * brand_3_sales


  set brand_buyer "man_1"
  let max_utility utility_man_1
  if max_utility <  utility_man_2 [set brand_buyer "man_2" set max_utility utility_man_2]
  if max_utility <  utility_man_3 [set brand_buyer "man_3" set max_utility utility_man_3]
  ;ifelse utility_man_1 > utility_threshold and elapsed = 1 [set brand_buyer 1] [set brand_buyer 0]

end
to repeat_purchase



  if last_brand_buyer != brand_buyer and elapsed = 1 [set habit (promo_coeff *   man_1_promotions) * (repeat_purchase_rate / 100)]
  if last_brand_buyer = brand_buyer and elapsed = 1 [set habit habit * (repeat_purchase_rate / 100)]
end
to brand_switcher

 if elapsed = 1 [set last_brand_buyer brand_buyer]

end
to ad_evaluate

;ifelse (remembered_exposures >= Total_Weighted_Frequency)[set active? true] [set active? false]
ifelse (man_1_media_one_grps > 0 )[set active_man_1? true] [set active_man_1? false]
ifelse (man_2_media_one_grps > 0 )[set active_man_2? true] [set active_man_2? false]
ifelse (man_3_media_one_grps > 0 )[set active_man_3? true] [set active_man_3? false]
;if  active? = true
;  [
;  ask link-neighbors with [active? = false]
;  [
;  if 1 > minimum_threshold [ set wom_exp  wom_weight ]]
;  ]
end
to recolor  ;; turtle
procedure
ifelse active? = true [set color 15 set size 1 ] [ifelse semiactive? = true [set color 4 set size 1 ] [set color 4 set size 1]]
end
to sec
set temp_sec (count consumers with [active? = true] / count consumers)
if ticks = 0 [set total_sec[]]
if ticks = 0 [set total_sec lput temp_sec total_sec]
if ticks > 0[ set total_sec lput (item (ticks - 1) total_sec + temp_sec) total_sec]
set report_sec ( item ticks  total_sec)
end
to neighborsd
  show (list who [node-id] of link-neighbors)
end
to seed
  ask n-of 20 consumers with [ycor < 5 and ycor > -5 and xcor < 5 and xcor > -5] [ set message_strength 10000 set color red]
end
to move
  ifelse (distancexy my_x my_y >= .1)
  [facexy my_x my_y fd .05]
  [if brand_buyer = 1 [lt random 360 fd 0.5]]
end
to color-links
ifelse ([color] of one-of both-ends = red)  [set color 17 set thickness 0]

                                              [set color 7 set thickness 0]
end
to-report get-node [id]
  report one-of consumers with [node-id = id]
end
to-report get-nodes [id]
  report [who] of one-of consumers with [node-id = id]
end
to-report seeder
 ; report min-one-of consumers  with [message_strength < minimum_threshold]  [ycor]
end
to my-setup-plots

  set-current-plot "Cross Media Exposure"
  set-plot-x-range 0 6
  set-plot-y-range 0 1
  set-histogram-num-bars 6

end
to do-cover-plots

set-current-plot "Market Shares"
```

```
    set-current-plot-pen "Producer 1 Sales"
    plot ((count consumers with [brand_buyer = "man_1"])/ count consumers)
    set-current-plot-pen "Producer 2 Sales"
    plot ((count consumers with [brand_buyer = "man_2"])/ count consumers)

       set-current-plot-pen "Producer 3 Sales"
    plot ((count consumers with [brand_buyer = "man_3"])/ count consumers)



      set-current-plot "advertising reach"
   plot count consumers with [ad_1_effective = 1]/ count consumers

   set-current-plot "Cross Media Exposure"
   histogram [cum_exp_count] of consumers


  set-current-plot "Purchase Cycle"
  histogram [current_pack] of consumers

    set-current-plot "utility"
  histogram [utility_man_1] of consumers  with [elapsed = 1]



end
to ad_assign_exposures
set remembered_exposures_man_1 remembered_exposures_man_1 + random-binomial (man_1_media_one_grps) (media_one_cons)
set remembered_exposures_man_1 remembered_exposures_man_1 * (1 - memory_decay)
set remembered_exposures_man_2 remembered_exposures_man_2 + random-binomial (man_2_media_one_grps) (media_one_cons)
set remembered_exposures_man_2 remembered_exposures_man_2 * (1 - memory_decay)


set remembered_exposures_man_3 remembered_exposures_man_3 + random-binomial (man_3_media_one_grps) (media_one_cons)
set remembered_exposures_man_3 remembered_exposures_man_3 * (1 - memory_decay)



end
to update_aggregates
          set brand_1_sales count consumers with [brand_buyer = "man_1"] / count consumers
          set brand_2_sales count consumers with [brand_buyer = "man_2"] / count consumers
          set brand_3_sales count consumers with [brand_buyer = "man_3"] / count consumers


end
to monitor_sales
set brand_1_sales_sum brand_1_sales_sum + count consumers with [brand_buyer = "man_1"]
set brand_2_sales_sum brand_2_sales_sum + count consumers with [brand_buyer = "man_2"]
set brand_3_sales_sum brand_3_sales_sum + count consumers with [brand_buyer = "man_3"]

end
;MANUFACTURERS
to select_strategy

  set chosen_strategy "promo"
  if ad_belief > promo_belief [set chosen_strategy "ad"]
  if coupons_belief > ad_belief  [set chosen_strategy "coupons"]

  if (ad_belief = promo_belief and promo_belief = coupons_belief and coupons_belief = ad_belief) [let c random 100 set chosen_strategy "ad" if c < 33 [set chosen_strategy "promo"] if c
> 66 [set chosen_strategy "coupons"]]

;  if (random 100 > 90) [let c random 100 set chosen_strategy "ad" if c < 33 [set chosen_strategy "promo"] if c > 66 [set chosen_strategy "coupons"]]
  if (ticks <= 20 and random 100 > 80) [set chosen_strategy "nothing"]

end
to-report random-binomial [n p]
   report length filter [? < p] n-values n [random-float 1]
end
to learn_effects

  show(shell:exec  "C:/Users/rstratton/Dropbox/coffee project/agent model/runsas man1.cmd")



    file-open "C:/Users/rstratton/Dropbox/coffee project/agent model/learning man1.txt"
  ;; Read in all the
 data in the file
  while [not file-at-end?]
  [
    ;; this reads a single line into a three-item list
    let items read-from-string (word "[" file-read-line "]")
    set man_1_imp1 item 0 items
    set man_1_imp2 item 1 items
    set man_1_imp3 item 2 items

    ]
  file-close


 show(shell:exec  "C:/Users/rstratton/Dropbox/coffee project/agent model/runsas man2.cmd")


    file-open "C:/Users/rstratton/Dropbox/coffee project/agent model/learning man2.txt"
```

```
 ;; Read in all the
data in the file
 while [not file-at-end?]
 [
   ;; this reads a single line into a three-item list
   let items read-from-string (word "[" file-read-line "]")
   set man_2_imp1 item 0 items
   set man_2_imp2 item 1 items
   set man_2_imp3 item 2 items

   ]
  file-close

ask manufacturer 2250 [ set ad_belief random 100 set promo_belief random 100 set coupons_belief random 100]
ask manufacturer 2251 [ set ad_belief man_2_imp1 set promo_belief man_2_imp2 set coupons_belief man_2_imp3]
  ask manufacturer 2252 [set ad_belief man_1_imp1 set promo_belief man_1_imp2 set coupons_belief man_1_imp3]

end
to deploy_policy

ask manufacturer 2250 [if chosen_strategy = "ad" [ set man_1_media_one_grps 100 set man_1_promotions 0 set man_1_coupons 0]
          if chosen_strategy = "promo" [ set man_1_media_one_grps 0 set man_1_promotions 1 set man_1_coupons 0]
          if chosen_strategy = "coupons" [ set man_1_media_one_grps 0 set man_1_promotions 0 set man_1_coupons 1]
          if chosen_strategy = "nothing" [ set man_1_media_one_grps 0 set man_1_promotions 0 set man_1_coupons 0]
          ]
ask manufacturer 2251 [if chosen_strategy = "ad" [ set man_2_media_one_grps 100 set man_2_promotions 0 set man_2_coupons 0]
          if chosen_strategy = "promo" [ set man_2_media_one_grps 0 set man_2_promotions 1 set man_2_coupons 0]
          if chosen_strategy = "coupons" [ set man_2_media_one_grps 0 set man_2_promotions 0 set man_2_coupons 1]
          if chosen_strategy = "nothing" [ set man_2_media_one_grps 0 set man_2_promotions 0 set man_2_coupons 0]
          ]
ask manufacturer 2252 [if chosen_strategy = "ad" [ set man_3_media_one_grps 100 set man_3_promotions 0 set man_3_coupons 0]
          if chosen_strategy = "promo" [ set man_3_media_one_grps 0 set man_3_promotions 1 set man_3_coupons 0]
          if chosen_strategy = "coupons" [ set man_3_media_one_grps 0 set man_3_promotions 0 set man_3_coupons 1]
          if chosen_strategy = "nothing" [ set man_3_media_one_grps 0 set man_3_promotions 0 set man_3_coupons 0]
          ]

end
; *** NetLogo 4.0 Code Example Copyright Notice ***
;
; (C) 2006 Uri Wilenorange.  This code may be freely copied, distributed,
; alte, or otherwise used by anyone for any legal purpose.
;
; THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS
; "AS IS" AND ANY _expRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT
; LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR
; A PARTICULAR PURPOSE ARE DISCLAIMED.  IN NO EVENT SHALL THE COPYRIGHT
; OWNERS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
; SPECIAL, EXEMPLARY, OR _consEQUENTIAL DAMAGES (INCLUDING, BUT NOT
; LIMITED TO,
PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE,
; DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY
; THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT
; (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE
; OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
;
; *** End of NetLogo 4.0 Code Example Copyright Notice ***
```

The code in this section relates to the implementations in Chapter 6.

```
 library("rjson")
json_file <- "C:/Users/Robert/Dropbox/marketing project/data/starbucks.json"

#json_file <-
"https://graph.facebook.com/164232989253/posts?limit=25000&access_token=AAACEdEose0cBAOeA2n5y98bKwEmAHzYfsoZBEpZCJhJF9f2k73GpPymAE9KTeAZAn5eAQhzqnHb8chyN
UZAuVBNi1VZBk8SEcM0Sahjy4sAZDZD&until=1299518462"

#json_data <- fromJSON(paste(readLines(json_file), collapse=""))

#print( json_data )
message<-json_data$data[[400]]$message
likes<-json_data$data[[400]]$likes$count
created<-json_data$data[[400]]$created_time

# download the file needed for authentication

#install.packages("rjson")


library("rjson")
#json_file <- "C:/Users/Robert/Dropbox/marketing project/data/starbucks.json"

json_file <-
"https://graph.facebook.com/164232989253/posts?limit=99000&access_token=AAACEdEose0cBAGbYcYgL2RyIqKQKNPg6FKFsZA7qxe5s9I80VJrb8bJ8WuxizHu0fjj1r2kIqIcjPLgY2fVz6vsSt
8ptPag0awXnc4QZDZD"

json_data <- fromJSON(paste(readLines(json_file), collapse=""))

print( json_data )
message<-""
likes<-""
created<-""
for (i in 1:446) {
message[i]<-""
likes[i]<-""
created[i]<-""
if (length(json_data$data[[i]]$message) > 0) message[i]<-json_data$data[[i]]$message
if (length(json_data$data[[i]]$likes$count) > 0) likes[i]<-json_data$data[[i]]$like$count
if (length(json_data$data[[i]]$created_time) > 0) created[i]<-json_data$data[[i]]$created_time

}

write.table(message, file="C:/Users/Robert/Dropbox/marketing project/data/message.csv",sep=",")
write.table(likes, file="C:/Users/Robert/Dropbox/marketing project/data/likes.csv",sep=",")
write.table(created, file="C:/Users/Robert/Dropbox/marketing project/data/created.csv",sep=",")


PROC IMPORT OUT= WORK.STARBUCKS
          DATAFILE= "C:\users\Rstratton\Dropbox\marketing project\data\processed data.csv"
          DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;
%let num_agents=10000;

data _null_;
call symput('num_agents_plus_one',(&num_agents + 1));

data agents;
do i= 1 to &num_agents;
social_pressure_threshold=1+floor(ranuni(0)*3) ;
random=ranuni(0);
threshold=i;
*if threshold >(&num_agents *.4) then threshold=(&num_agents *.4);
p_Intercept=47.42425;
p_trend=43.51278;
p_latte=569.83071;
p_frappucino=1040.13978;
p_coffee=-117.03499;
p_textlength=-3.69186;
p_qmark=-204.50863;
p_exmark=472.77329;
p_weekday_1=878.96516;
p_weekday_2=619.25915;
p_weekday_7=370.88721;
p_weekday_4=191.32682;
p_weekday_5=522.24722;
p_weekday_6=607.39175;
p_days_since_last=-38.02641;
p_hour=-79.89834;
output;
```

```
end;

data agents;
retain agent_id 0;
set agents;
agent_id= agent_id +1;
agent_contact_1=1+floor(ranuni(0)*&num_agents);
agent_contact_2=1+floor(ranuni(0)*&num_agents);
agent_contact_3=1+floor(ranuni(0)*&num_agents);
agent_contact_4=1+floor(ranuni(0)*&num_agents);
run;
%macro assign_agents;
data agents;
set agents;
%mend;

%assign_agents;

run;
data brand_page;
retain time 0;
set starbucks;
time=time + 1;

data catch;
data catch_viral;
data catch_natural;

%macro time_loop;
%do i= 1 %to 115;
proc sql;
create table consider as select * from brand_page a, agents b where a.time=&i;;

data calculate;
length agent_contacted_1-agent_contacted_&num_agents $1.;
retain agent_contacted_1-agent_contacted_&num_agents "0";
array agent_contacted_(&num_agents_plus_one) ;
set consider;

natural_like=0;
viral_like=0;

my_contacts = (agent_contacted_(agent_id));

latent_like=p_Intercept+
p_trend*trend+
p_latte*latte+
p_frappucino*frappucino+
p_coffee*coffee+
p_textlength*textlength+
p_qmark*qmark+
p_exmark*exmark+
p_weekday_1*weekday_1+
p_weekday_2*weekday_2+
p_weekday_7*weekday_7+
p_weekday_4*weekday_4+
p_weekday_5*weekday_5+
p_weekday_6*weekday_6+
p_days_since_last*days_since_last+
p_hour*hour;
model_like=0;
if latent_like > threshold then do;model_like=1; natural_like=1;end;

if (my_contacts="4" or my_contacts="5" or my_contacts="6")  then do; model_like=1; viral_like=1; end;

if model_like =1 and agent_contacted_(agent_contact_1) ="7" then agent_contacted_(agent_contact_1) ="8";
if model_like =1 and agent_contacted_(agent_contact_1) ="6" then agent_contacted_(agent_contact_1) ="7";
if model_like =1 and agent_contacted_(agent_contact_1) ="5" then agent_contacted_(agent_contact_1) ="6";
if model_like =1 and agent_contacted_(agent_contact_1) ="4" then agent_contacted_(agent_contact_1) ="5";
if model_like =1 and agent_contacted_(agent_contact_1) ="3" then agent_contacted_(agent_contact_1) ="4";
if model_like =1 and agent_contacted_(agent_contact_1) ="2" then agent_contacted_(agent_contact_1) ="3";
if model_like =1 and agent_contacted_(agent_contact_1) ="1" then agent_contacted_(agent_contact_1) ="2";
if model_like =1 and agent_contacted_(agent_contact_1) ="0" then agent_contacted_(agent_contact_1) ="1";

if model_like =1 and agent_contacted_(agent_contact_2) ="7" then agent_contacted_(agent_contact_2) ="8";
if model_like =1 and agent_contacted_(agent_contact_2) ="6" then agent_contacted_(agent_contact_2) ="7";
if model_like =1 and agent_contacted_(agent_contact_2) ="5" then agent_contacted_(agent_contact_2) ="6";
if model_like =1 and agent_contacted_(agent_contact_2) ="4" then agent_contacted_(agent_contact_2) ="5";
if model_like =1 and agent_contacted_(agent_contact_2) ="3" then agent_contacted_(agent_contact_2) ="4";
if model_like =1 and agent_contacted_(agent_contact_2) ="2" then agent_contacted_(agent_contact_2) ="3";
if model_like =1 and agent_contacted_(agent_contact_2) ="1" then agent_contacted_(agent_contact_2) ="2";
if model_like =1 and agent_contacted_(agent_contact_2) ="0" then agent_contacted_(agent_contact_2) ="1";
```

```
if model_like =1 and agent_contacted_(agent_contact_3) ="7" then agent_contacted_(agent_contact_3) ="8";
if model_like =1 and agent_contacted_(agent_contact_3) ="6" then agent_contacted_(agent_contact_3) ="7";
if model_like =1 and agent_contacted_(agent_contact_3) ="5" then agent_contacted_(agent_contact_3) ="6";
if model_like =1 and agent_contacted_(agent_contact_3) ="4" then agent_contacted_(agent_contact_3) ="5";
if model_like =1 and agent_contacted_(agent_contact_3) ="3" then agent_contacted_(agent_contact_3) ="4";
if model_like =1 and agent_contacted_(agent_contact_3) ="2" then agent_contacted_(agent_contact_3) ="3";
if model_like =1 and agent_contacted_(agent_contact_3) ="1" then agent_contacted_(agent_contact_3) ="2";
if model_like =1 and agent_contacted_(agent_contact_3) ="0" then agent_contacted_(agent_contact_3) ="1";

if model_like =1 and agent_contacted_(agent_contact_4) ="7" then agent_contacted_(agent_contact_4) ="8";
if model_like =1 and agent_contacted_(agent_contact_4) ="6" then agent_contacted_(agent_contact_4) ="7";
if model_like =1 and agent_contacted_(agent_contact_4) ="5" then agent_contacted_(agent_contact_4) ="6";
if model_like =1 and agent_contacted_(agent_contact_4) ="4" then agent_contacted_(agent_contact_4) ="5";
if model_like =1 and agent_contacted_(agent_contact_4) ="3" then agent_contacted_(agent_contact_4) ="4";
if model_like =1 and agent_contacted_(agent_contact_4) ="2" then agent_contacted_(agent_contact_4) ="3";
if model_like =1 and agent_contacted_(agent_contact_4) ="1" then agent_contacted_(agent_contact_4) ="2";
if model_like =1 and agent_contacted_(agent_contact_4) ="0" then agent_contacted_(agent_contact_4) ="1";

proc sql;
create table output as select sum(model_like) as model_likes, date as date from calculate group by date;

proc sql;
create table natural_output as select sum(natural_like) as natural_likes, date as date from calculate group by date;

proc sql;
create table viral_output as select sum(viral_like) as viral_likes, date as date from calculate group by date;

data catch;
set catch output;
run;

data catch_viral;
set catch_viral viral_output;

data catch_natural;
set catch_natural natural_output;


data calculate_&i;
set calculate;
run;
%end;
%mend;

%time_loop;
run;




proc sql;
create table joined as select * from starbucks a, catch b where a.date=b.date;

proc reg data=joined;
model likes=model_likes;
run;

proc gplot;
plot model_likes*date likes*date/overlay;
SYMBOL1 value=dot  interpol=join line=1 color=red width=1;
SYMBOL2 value=dot  interpol=join line=1 color=green width=1;
run;

/*
data agents;
do i= 1 to &num_agents;
social_pressure_threshold=1+floor(ranuni(0)*4) ;
random=ranuni(0);
threshold=i;
if threshold >20 then threshold=20;
p_Intercept=3347.14064;
p_latte=1188.74917;
p_frappucino=1417.46666;
p_coffee=-77.8874;
p_textlength=-11.12057;
p_qmark=-175.0379;
p_exmark=388.57785;
p_weekday_1=726.74736;
p_weekday_2=851.76293;
p_weekday_7=-918.95152;
p_weekday_4=270.0394;
```

```
p_weekday_5=303.64466;
p_weekday_6=-9.26041;
p_days_since_last=-28.50301;
p_hour=-74.61624;
output;
end;

if (my_contacts="1" and social_pressure_threshold =1)
or (my_contacts="2" and (social_pressure_threshold =1 or social_pressure_threshold=2))
or (my_contacts="3" and (social_pressure_threshold =1 or social_pressure_threshold=2 or social_pressure_threshold=3))
or (my_contacts="4" and (social_pressure_threshold =1 or social_pressure_threshold=2 or social_pressure_threshold=3 or social_pressure_threshold=4))
then model_like=1;
*/


PROC IMPORT OUT= WORK.STARBUCKS
        DATAFILE= "C:\users\Rstratton\Dropbox\marketing project\data\processed data.csv"
        DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

data sb;
set starbucks;
sq_latte=latte*latte;
sq_frappucino=frappucino*frappucino;
sq_coffee=coffee*coffee;
sq_textlength=textlength*textlength;
sq_qmark=qmark*qmark;
sq_exmark=exmark*exmark;
sq_weekday_1=weekday_1*weekday_1;
sq_weekday_2=weekday_2*weekday_2;
sq_weekday_7=weekday_7*weekday_7;
sq_weekday_4=weekday_4*weekday_4;
sq_weekday_5=weekday_5*weekday_5;
sq_weekday_6=weekday_6*weekday_6;
sq_days_since_last=days_since_last*days_since_last;
sq_hour=hour*hour;
log_latte=log(latte+1);
log_frappucino=log(frappucino+1);
log_coffee=log(coffee+1);
log_textlength=log(textlength+1);
log_qmark=log(qmark+1);
log_exmark=log(exmark+1);
log_weekday_1=log(weekday_1+1);
log_weekday_2=log(weekday_2+1);
log_weekday_7=log(weekday_7+1);
log_weekday_4=log(weekday_4+1);
log_weekday_5=log(weekday_5+1);
log_weekday_6=log(weekday_6+1);
log_days_since_last=log(days_since_last+1);
log_hour=log(hour+1);
wom=0;
if likes > 2000 then wom=likes-2000;
run;


proc reg data=sb;
model likes= trend latte frappucino coffee textlength qmark exmark weekday_1 weekday_2 weekday_7 weekday_4 weekday_5 weekday_6 days_since_last hour
;
output out=a p=pred;
run;
quit;

proc gplot;
plot pred*date loglikes*date/overlay;
SYMBOL1 value=dot  interpol=join line=1 color=red width=1;
SYMBOL2 value=dot  interpol=join line=1 color=green width=1;
run;


proc gplot;
plot pred*date loglikes*date/overlay;
SYMBOL1 value=dot  interpol=join line=1 color=red width=1;
SYMBOL2 value=dot  interpol=join line=1 color=green width=1;
run;


proc gplot;
plot pred*date loglikes*date/overlay;
SYMBOL1 value=dot  interpol=join line=1 color=red width=1;
SYMBOL2 value=dot  interpol=join line=1 color=green width=1;
```

```
run;


proc reg data=sb;
model loglikes=

trend latte frappucino coffee textlength qmark exmark weekday_1 weekday_2 weekday_7 weekday_4 weekday_5 weekday_6 days_since_last hour
log_latte
log_frappucino
log_coffee
log_textlength
log_qmark
log_exmark
log_weekday_1
log_weekday_2
log_weekday_7
log_weekday_4
log_weekday_5
log_weekday_6
log_days_since_last
log_hour
;
quit;
run;
proc reg data=starbucks;
model likes=trend latte frappucino coffee textlength qmark exmark weekday_1 weekday_2 weekday_7 weekday_4 weekday_5 weekday_6 days_since_last hour;
run;

proc univariate data=starbucks;
var likes;
histogram likes;
run;
```

The code in this section relates to the implementations in Chapter 7.

```
install.packages("msm")
library("msm")
mym<- read.csv("C:/Users/rstratton/Dropbox/autonomous agent project/analysis/sas_output_data.csv")
twoway4.q <- rbind(c(.5, .11, .11), c(.9,.1,.11 ), c(.9,.1,.1 ))
# I want to loop around from HERE for each value of the field ID:
my.msm <- msm (  vol ~ month, subject=house, data = mym, twoway4.q ,
                  hmodel = list(hmmUnif(0, 2), hmmUnif(2, 3), hmmUnif(2, 11)),
                        covariates=~ christmas + easter,
          )
pmatrix.msm(my.msm, covariates = list (1, 0),
         ci=c("none","normal","bootstrap"), cl=0.95,
        )
pmatrix.msm(my.msm, covariates = list (0, 1),
         ci=c("none","normal","bootstrap"), cl=0.95,
)

pmatrix.msm(my.msm, covariates = list (0, 0),
         ci=c("none","normal","bootstrap"), cl=0.95,
)
```

SAS Code

```
/*INITIALISE*/

libname livet "C:\users\rstratton\Dropbox\me\livet\SAS data";
run;
data t;
set livet.analysis_file;
run;
proc sql;
create table purchases as select house, relweek, sum(decision) as decisions from t  group by house,relweek;
proc sql;
create table agents as select house, count(*) from purchases where relweek < 20 group by house;

proc sql;
create table t as select * from t, agents where t.house=agents.house;
*INITIALISE MEMORIES;
proc sql;
create table last_brand as select house, at4, relweek, count(*) from t where relweek < 20 and decision = 1 group by house, relweek, at4 ;
proc sort data=last_brand;
by house relweek;
data last_brand2;
set last_brand;
by house;
if last.house then output;
run;
data last;
set last_brand2;
last_brand=at4;
keep last_brand house;

proc sql;
create table t as select * from t, last where t.house=last.house;
*INITIALISE DISCUSSED PREFERENCES;
proc sql;
create table brands as select at4, count(*) from t group by at4;

data brands;
set brands;
brand=at4;

proc sql;
create table sum_purchases as select at4, sum(decision) as decisions from t where relweek < 20 group by at4;

proc sql;
select  sum(decision)into :grand_total from t where relweek < 20 ;


proc sort data=sum_purchases;
by at4;
proc sort data=brands;
by at4;
data initialise_prefs;
merge brands(in=a) sum_purchases(in=b) ;
by at4;
if a;
if decisions=. then decisions=0;
run;

data initialise_prefs;
set initialise_prefs;
discussed_pct=decisions/&grand_total;
```

```
keep at4 discussed_pct;
run;
proc sql;
create table t as select * from initialise_prefs a,  t b where a.at4=b.at4;

data t;
set t;
if last_brand=at4 then last_purchase=1;
if last_brand ^=at4 then last_purchase=0;
run;

%macro simulate_loop;


data simulate;
set t;
*THIS SWITCHES IT TO EMERGENT MODE;
*where relweek  = &my_relweek;
where relweek  = 111;

run;

proc sql;
create table simulator as select * from simulate a, store_parms b where a.house=b.house;

run;


data score;
set simulator;
p=sum(
(my_p_discussed_pct*discussed_pct),
(my_p_last_purchase=last_purchase),

(my_p_multibuy * multibuy ),
(my_p_tpr_discount * tpr_discount ),
(my_p_advertising * advertising ),
(my_p_price_penalty * price_penalty ),
(my_p_vol_penalty * vol_penalty ),
(my_p_health_dummy * health_dummy ),
(my_p_branded_dummy * branded_dummy ),
(my_p_olive_dummy * olive_dummy ),
(my_p_lowfat_dummy * lowfat_dummy ),
(my_p_cholesterol_dummy * cholesterol_dummy ),
(my_p_spreadable_dummy *spreadable_dummy));
keep house date relweek shopcode available_brand_product actual_purchase event_count p;
run;

proc sort data=score;
by house event_count p;

data my_purchase;
set score;
by house event_count;
if last.event_count then output;
r=&my_relweek;

run;

data simulate_purchase_store;
set simulate_purchase_store my_purchase;

%mend;
run;

%macro simulate_discuss_purchases;

data purchase_store;
length at4 $100.;
set simulate_purchase_store;
at4=scan(available_brand_product,1,"|");
run;

proc sql;
 create table discussed_parms as select at4, count(*) as count from purchase_store group by at4 ;

proc sort data=discussed_parms;
 by at4;

data discussed_prefs;
merge brands(in=a) discussed_parms(in=b) ;
by at4;
```

```
if a;
if count=. then count=0;
run;

proc sql;
select sum(count) into :sum_prefs from discussed_prefs;

data discussed_prefs;
set discussed_prefs;
new_discussed_pct=count/&sum_prefs;
at4=brand;
keep at4 new_discussed_pct;
run;

data discussed;
set discussed_prefs;
%do i= 1 %to &total_house_count;
%do j=(&my_relweek + 1) %to 151;
relweek=&j;
house=&&total_house&i;
output;
%end;
%end;
run;

proc sort data=discussed;
by house relweek at4;
run;

proc sort data=t;
by house relweek at4;

data t;
merge t (in=a) discussed (in=b);
by house relweek at4;
if a;
run;


data t;
set t;
if new_discussed_pct ^=. then discussed_pct = new_discussed_pct;

data d3;
set t;
run;
%mend;

%macro simulate_update_memories;

data my_memories;
set purchase_store;

proc sort data=my_memories;
by house relweek;

data my_memories;
set my_memories;
my_new_memory=at4;
by house;
if last.house then output;

data memories;
set my_memories;
%do j=(&my_relweek + 1)  %to &maxweeks;
relweek=&j;
output;
%end;

run;


data t;
set t;
joiner=left(trim(house))||"_"||left(trim(relweek))||"_"||left(trim(at4));

data memories;
set memories;
joiner=left(trim(house))||"_"||left(trim(relweek))||"_"||left(trim(at4));
keep joiner my_new_memory;

proc sort data=memories;
```

```
by joiner;
run;

proc sort data=t;
by joiner;

data t;
merge t (in=a) memories (in=b);
by joiner;
if a;
run;

data t;
set t;
if my_new_memory ^="" then last_purchase=0;
if my_new_memory =at4 then last_purchase=1;

run;

data d4;
set t;
run;

%mend;

%macro initialise;

data t;
set livet.analysis_file;
run;

proc sql;
create table purchases as select house, relweek, sum(decision) as decisions from t  group by house,relweek;

proc sql;
create table agents as select house, count(*) from purchases where relweek < 20 group by house;

proc sql;
create table t as select * from t, agents where t.house=agents.house;

*INITIALISE MEMORIES;

proc sql;
create table last_brand as select house, at4, relweek, count(*) from t where relweek < 20 and decision = 1 group by house, relweek, at4 ;


proc sort data=last_brand;
by house relweek;


data last_brand2;
set last_brand;
by house;
if last.house then output;
run;

data last;
set last_brand2;
last_brand=at4;
keep last_brand house;

proc sql;
create table t as select * from t, last where t.house=last.house;


*INITIALISE DISCUSSED PREFERENCES;

proc sql;
create table brands as select at4, count(*) from t group by at4;

data brands;
set brands;
brand=at4;


proc sql;
create table sum_purchases as select at4, sum(decision) as decisions from t where relweek < 20 group by at4;


proc sql;
select  sum(decision)into :grand_total from t where relweek < 20 ;
```

```
proc sort data=sum_purchases;
by at4;

proc sort data=brands;
by at4;

data initialise_prefs;
merge brands(in=a) sum_purchases(in=b) ;
by at4;
if a;
if decisions=. then decisions=0;
run;


data initialise_prefs;
set initialise_prefs;
discussed_pct=decisions/&grand_total;

keep at4 discussed_pct;
run;

proc sql;
create table t as select * from t a,  initialise_prefs b where a.at4=b.at4;

data t;
set t;
if last_brand=at4 then last_purchase=1;
if last_brand ^=at4 then last_purchase=0;
run;

data d1;
set t;
run;
%mend;

%macro loop;
%do i= 1 %to &house_count;

data analysis;
set t;
where house=&&house&i and relweek<=&my_relweek;

proc mdc data=analysis outest=c noprint;

model decision = vol_penalty
discussed_pct
last_purchase
price_penalty
olive_dummy
cholesterol_dummy
lowfat_dummy
 health_dummy

 branded_dummy
spreadable_dummy
advertising
tpr_discount
 multibuy

/
type=clogit   choice=(mode )
;

id event_count;
output out=probdata pred=p;

*ods output  ParameterEstimates = parms;

run;

data my_parms;
set c;
my_p_discussed_pct=discussed_pct;
my_p_last_purchase=last_purchase;
my_p_multibuy =multibuy ;
my_p_tpr_discount = tpr_discount ;
my_p_advertising = advertising ;
my_p_price_penalty = price_penalty ;
my_p_vol_penalty = vol_penalty ;
my_p_health_dummy = health_dummy ;
my_p_branded_dummy = branded_dummy ;
```

```
my_p_spreadable_dummy = spreadable_dummy;
my_p_olive_dummy = olive_dummy;
my_p_lowfat_dummy = lowfat_dummy;
my_p_cholesterol_dummy = cholesterol_dummy;

drop
 tpr_discount
 advertising
price_penalty
 vol_penalty
 health_dummy
 branded_dummy
 spreadable_dummy
olive_dummy
lowfat_dummy
cholesterol_dummy
 multibuy

 discussed_pct
last_purchase
;

proc sql;
create table scored as
select * from analysis, my_parms;
quit;


data my_probdata;
set scored;
p=sum(
(my_p_discussed_pct*discussed_pct),
(my_p_last_purchase=last_purchase),

(my_p_multibuy * multibuy ),
(my_p_tpr_discount * tpr_discount ),
(my_p_advertising * advertising ),
(my_p_price_penalty * price_penalty ),
(my_p_vol_penalty * vol_penalty ),
(my_p_health_dummy * health_dummy ),
(my_p_branded_dummy * branded_dummy ),
(my_p_olive_dummy * olive_dummy ),
(my_p_lowfat_dummy * lowfat_dummy ),
(my_p_cholesterol_dummy * cholesterol_dummy ),
(my_p_spreadable_dummy *spreadable_dummy));
keep house date relweek shopcode available_brand_product actual_purchase event_count p;
run;

proc sort data=my_probdata;
by house event_count p;


data my_parms;
set my_parms;
house=&&house&i;
relweek= &my_relweek;

data my_purchase;
set my_probdata;
by house event_count;
if last.event_count then output;


data purchase_store;
set purchase_store my_purchase;

data store_parms;
set store_parms my_parms;

data store_c;
set store_c c;

run;
%end;
%mend;
run;

%macro discuss_purchases;

data purchase_store;
length at4 $100.;
set purchase_store;
```

```
at4=scan(actual_purchase,1,"|");
run;

proc sql;
 create table discussed_parms as select at4, count(*) as count from purchase_store group by at4 ;

proc sort data=discussed_parms;
 by at4;

data discussed_prefs;
merge brands(in=a) discussed_parms(in=b) ;
by at4;
if a;
if count=. then count=0;
run;

proc sql;
select sum(count) into :sum_prefs from discussed_prefs;

data discussed_prefs;
set discussed_prefs;
new_discussed_pct=count/&sum_prefs;
at4=brand;
keep at4 new_discussed_pct;
run;

data discussed;
set discussed_prefs;
%do i= 1 %to &total_house_count;
%do j=(&my_relweek + 1) %to &maxweeks;
relweek=&j;
house=&&total_house&i;
output;
%end;
%end;
run;

proc sort data=discussed;
by house relweek at4;
run;

proc sort data=t;
by house relweek at4;

data t;
merge t (in=a) discussed (in=b);
by house relweek at4;
if a;
run;

data t;
set t;
if new_discussed_pct ^=. then discussed_pct = new_discussed_pct;
*THIS LINE SWITCHES OFF DISCUSSION;
*discussed_pct=0;

run;
%mend;

%macro update_memories;

data my_memories;
set purchase_store;

proc sort data=my_memories;
by house relweek;

data my_memories;
set my_memories;
my_new_memory=at4;
by house;
if last.house then output;

data memories;
set my_memories;
%do j=(&my_relweek + 1)  %to &maxweeks;
relweek=&j;
output;
%end;

run;
```

```
data t;
set t;
joiner=left(trim(house))||"_"||left(trim(relweek))||"_"||left(trim(at4));

data memories;
set memories;
joiner=left(trim(house))||"_"||left(trim(relweek))||"_"||left(trim(at4));
keep joiner my_new_memory;

proc sort data=memories;
by joiner;
run;

proc sort data=t;
by joiner;

data t;
merge t (in=a) memories (in=b);
by joiner;
if a;
run;

data t;
set t;
if my_new_memory ^="" then last_purchase=0;
if my_new_memory =at4 then last_purchase=1;

data d4;
set t;
run;

%mend;
run;
%macro simulate_loop;
data simulate;
set t;
*THIS SWITCHES IT TO EMERGENT MODE;
*where relweek  = &my_relweek;
where relweek  = 111;
run;
proc sql;
create table simulator as select * from simulate a, store_parms b where a.house=b.house;
data score;
set simulator;
p=sum(
(my_p_discussed_pct*discussed_pct),
(my_p_last_purchase=last_purchase),
(my_p_multibuy * multibuy ),
(my_p_tpr_discount * tpr_discount ),
(my_p_advertising * advertising ),
(my_p_price_penalty * price_penalty ),
(my_p_vol_penalty * vol_penalty ),
(my_p_health_dummy * health_dummy ),
(my_p_branded_dummy * branded_dummy ),
(my_p_olive_dummy * olive_dummy ),
(my_p_lowfat_dummy * lowfat_dummy ),
(my_p_cholesterol_dummy * cholesterol_dummy ),
(my_p_spreadable_dummy *spreadable_dummy));
keep house date relweek shopcode available_brand_product actual_purchase event_count p;
run;

proc sort data=score;
by house event_count p;

data my_purchase;
set score;
by house event_count;
if last.event_count then output;
r=&my_relweek;

data simulate_purchase_store;
set simulate_purchase_store my_purchase;

%mend;
run;

%macro simulate_discuss_purchases;
data purchase_store;
length at4 $100.;
set simulate_purchase_store;
at4=scan(available_brand_product,1,"|");
run;
```

```
proc sql;
 create table discussed_parms as select at4, count(*) as count from purchase_store group by at4 ;

proc sort data=discussed_parms;
 by at4;

data discussed_prefs;
merge brands(in=a) discussed_parms(in=b) ;
by at4;
if a;
if count=. then count=0;
run;

proc sql;
select sum(count) into :sum_prefs from discussed_prefs;

data discussed_prefs;
set discussed_prefs;
new_discussed_pct=count/&sum_prefs;
at4=brand;
keep at4 new_discussed_pct;
run;

data discussed;
set discussed_prefs;
%do i= 1 %to &total_house_count;
%do j=(&my_relweek + 1) %to 151;
relweek=&j;
house=&&total_house&i;
output;
%end;
%end;
run;

proc sort data=discussed;
by house relweek at4;
run;

proc sort data=t;
by house relweek at4;

data t;
merge t (in=a) discussed (in=b);
by house relweek at4;
if a;
run;

data t;
set t;
if new_discussed_pct ^=. then discussed_pct = new_discussed_pct;

data d3;
set t;
run;
%mend;

%macro simulate_update_memories;

data my_memories;
set purchase_store;

proc sort data=my_memories;
by house relweek;

data my_memories;
set my_memories;
my_new_memory=at4;
by house;
if last.house then output;

data memories;
set my_memories;
%do j=(&my_relweek + 1)  %to &maxweeks;
relweek=&j;
output;
%end;

data t;
set t;
joiner=left(trim(house))||"_"||left(trim(relweek))||"_"||left(trim(at4));
```

```
data memories;
set memories;
joiner=left(trim(house))||"_"||left(trim(relweek))||"_"||left(trim(at4));
keep joiner my_new_memory;

proc sort data=memories;
by joiner;
run;

proc sort data=t;
by joiner;

data t;
merge t (in=a) memories (in=b);
by joiner;
if a;
run;

data t;
set t;
if my_new_memory ^="" then last_purchase=0;
if my_new_memory =at4 then last_purchase=1;

run;

data d4;
set t;
run;

%mend;
```

Bibliography

[1] Magid Abraham and Leonard Lodish. Promoter: An automated promotion evaluation system. *Marketing Science*, 6(2):101–123, 1987.

[2] GM Allenby and Peter Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89, 1998.

[3] Frederic Amblard, Pierre Bommel, and Juliette Rouchier. *Assessment and Validation of Multi-Agent Models*, pages 93–114. GEMAS Studies in Social Analysis. Bardwell, 2007.

[4] Henrik Amilon. Estimation of an adaptive stock market model with heterogeneous agents. *Journal of Empirical Finance*, 15(2):342–362, March 2008.

[5] Arnold E Amstutz. *Computer Simulation of Competitive Market Response*, volume 1. The MIT Press, 1 edition, 1970.

[6] John Anderson, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1050, 2004.

[7] Paul F Anderson. Marketing, strategic planning and the theory of the firm. *Journal of Marketing*, 46(2):pp. 15–26, 1982.

[8] Giulia Andrighetto and M Campennì. The complex loop of norm emergence: A simulation model. In *Simulating Interacting Agents and Social Phenomena*, volume 7, pages 19–35. Springer Japan, 2010.

[9] Gianluca Antonini, Michel Bierlaire, and Mats Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006.

[10] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[11] Robert Axelrod and Leigh Tesfatsion. A guide for newcomers to agent-based modeling in the social sciences. Staff General Research Papers 12515, Iowa State University, Department of Economics, January 2006.

[12] Robert Axtell. Why agents? On the varied motivations for agent computing in the social sciences. In *Working Paper 17, Center on Social and Economic Dynamics, Brookings Institution*, volume 17, 2000.

[13] Alan Baker. Simulation-based definitions of emergence. *Journal of Artificial Societies and Social Simulation*, 13(1):9, 2010.

[14] BM Baker and MA Ayechew. A genetic algorithm for the vehicle routing problem. *Computers & Operations Research*, 30:787–800, 2003.

[15] Chris Baker, Joshua Tenenbaum, and Rebecca Saxe. Bayesian models of human action understanding. *Advances in Neural Inforrmation Processing Systems*, 18:99, 2006.

[16] Santiago Balseiro and J Feldman. Yield optimization of display advertising with ad exchange. *Proceedings of the 12th ACM conference on Electronic commerce*, 2011.

[17] Stefania Bandini, Sara Manzoni, and Giuseppe Vizzari. Agent based modeling and simulation: an informatics perspective. *Journal of Artificial Societies and Social Simulation*, 12(4):4, 2009.

[18] Omar Baqueiro and YJ Wang. Integrating data mining and agent based modeling and simulation. In Petra Perner, editor, *Advances in Data Mining. Applications and Theoretical Aspects*, volume 5633 of *Lecture Notes in Computer Science*, pages 220–231. Springer Berlin / Heidelberg, 2009.

[19] Frank M Bass, Moshe M Givon, Manohar U Kalwani, David Reibstein, and Gordon P Wright. An investigation into the order of the brand choice process. *Marketing Science*, 3(4):267–287, 1984.

[20] Mark Bedau. Downward causation and autonomy in weak emergence. *Principia*, 6:5–50, 2003.

[21] Moshe Ben-Akiva and Michel Bierlaire. Discrete choice methods and their applications to short term travel decisions. *Handbook of Transportation Science*, pages 5—-33, 1999.

[22] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, July 1995.

[23] Bruce Biegel. The current view and outlook for the future of marketing automation. *Journal of Direct, Data and Digital Marketing Practice*, 10(3):201–213, January 2009.

[24] Andreas Birk. Boosting cooperation by evolving trust. *Applied Artificial Intelligence*, 14:769–784, 2000.

[25] Christian Borgs, Jennifer Chayes, Nicole Immorlica, Kamal Jain, Omid Etesami, and Mohammad Mahdian. Dynamics of bid optimization in online advertisement auctions. *Proceedings of the 16th International Conference on World Wide Web - WWW '07*, page 531, 2007.

[26] William F Brewer and Bruce L Lambert. The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process. *Philosophy of Science*, 68, 2001.

[27] David Broadhurst, Royston Goodacre, Alun Jones, Jem J Rowland, and Douglas B Kell. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta*, 348(1-3):71–86, August 1997.

[28] Rodney A Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–160, 1991.

[29] Joanna J Bryson. Action selection and individuation in agent based modelling. *The Proceedings of Agent 2003: Challenges of Social Simulation*, pages 1–18, 2003.

[30] Randolph Bucklin, Donald Lehmann, and John Little. From decision support to decision automation: A 2020 vision. *Marketing Letters*, 9(3):235–246, 1998.

[31] Kenneth Burnham and David Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, July 2002.

[32] Benoit Calvez and Guillaume Hutzler. Automatic tuning of agent-based models using genetic algorithms. In Jaime S Sichman and Luis Antunes, editors, *Multi-Agent-Based Simulation VI*, volume 3891, chapter 4, pages 41–57. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[33] Cristiano Castelfranchi, Rosaria Conte, and Mario Paolucci. Normative reputation and the costs of compliance. *Journal of Artificial Societies and Social Simulation*, vol. 1, no, 1998.

[34] Jennifer L Castle. Evaluating PcGets and RETINA as automatic model selection algorithms. *Oxford Bulletin of Economics and Statistics*, 67:837–880, 2005.

[35] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and PK Gummadi. Measuring user influence in Twitter: the million follower fallacy. In *ICWSM*, 2010.

[36] Shu-Heng Chen, CL Chang, and YR Du. Agent-based economic models and econometrics. *Knowledge Engineering Review*, 27(1):187–219, 2012.

[37] Ding-An Chiang, Yi-Fan Wang, Shao-Lun Lee, and Cheng-Jung Lin. Goal-oriented sequential pattern for network banking churn analysis. *Expert Systems with Applications*, 25(3):293–302, 2003.

[38] Hui-Qing Chong, Ah-Hwee Tan, and Gee-Wah Ng. Integrated cognitive architectures: a survey. *Artificial Intelligence Review*, 28(2):103–130, August 2007.

[39] Pasquale Cirillo and Mauro Gallegati. The empirical validation of an agent-based model. *Eastern Economic Journal*, 38(4):525–547, February 2012.

[40] Peter Clark and Stan Matwin. Learning domain theories using abstract background knowledge. *Machine Learning: ECML-93*, pages 1–18, 1993.

[41] Rosaria Conte. From simulation to theory (and backward). In Flaminio Squazzoni, editor, *Epistemological Aspects of Computer Simulation in the Social Sciences*, pages 29–47. Springer-Verlag, Berlin, Heidelberg, 2009.

[42] Rosaria Conte, Bruce Edmonds, Scott Moss, and R Keith Sawyer. Sociology and social theory in agent based social simulation: A symposium. *Computational Mathematical Organization Theory*, 7(3):183–205, 2001.

[43] Lee G Cooper and Masao Nakanishi. *Market-share Analysis*. Number [1] in International series in quantitative marketing. Kluwer, 2. print edition, 1990.

[44] Jason Davis, Kathleen Eisenhardt, and Christopher Bingham. Developing theory through simulation methods. *Academy of Management Review*, 32(2):21, November 2007.

[45] Scott de Marchi. *Computational and Mathematical Modeling in the Social Sciences*. Cambridge University Press, 2005.

[46] Lisette de Vries, Sonja Gensler, and Peter S.H. Leeflang. Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing. *Journal of Interactive Marketing*, 26(2):83–91, May 2012.

[47] Sebastiano Delre, Wander Jager, Tammo Bijmolt, and Marco Janssen. Targeting and timing promotional activities: An agent-based model for the takeoff of new products. *Journal of Business Research*, 60(8):826–835, 2007.

[48] Steven K Donoho and Larry A Rendell. Rerepresenting and restructuring domain theories: A constructive induction approach. *arXiv preprint cs/9504101*, 2:411–446, 1995.

[49] Alexis Drogoul, Diane Vanbergue, and Thomas Meurisse. Multi-agent based simulation: Where are the agents? In Jaime Simão Sichman, Françcois Bousquet, and Paul Davidsson, editors, *Multi-agent-based simulation II*, volume 2581 of *Lecture Notes in Computer Science*, pages 43–49. Springer Berlin / Heidelberg, 2003.

[50] Saso Dzeroski and Ljupco Todorovski, editors. *Computational Discovery of Scientific Knowledge, Introduction, Techniques, and Applications in Environmental and Life Sciences*, volume 4660 of *Lecture Notes in Computer Science*. Springer, 2007.

[51] Robert East, Malcolm Wright, and Marc Vanheule. *Consumer Behaviour, Applications in Marketing*. Sage, 2009.

[52] Bruce Edmonds and Joanna J Bryson. The Insufficiency of Formal Design Methods ” The Necessity of an Experimental Approach - for the Understanding and Control of Complex MAS. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS ’04, pages 938–945, Washington, DC, USA, 2004. IEEE Computer Society.

[53] Bruce Edmonds, Emma Norling, and David Hales. Towards the evolution of social structure. *Computational Mathematical Organization Theory*, 15(2):78–94, 2009.

[54] Eric Eisenstein and Leonard Lodish. Marketing decision support and intelligent systems: precisely worthwhile or vaguely worthless? *Handbook of Marketing*, pages 436–454, 2002.

[55] Corinna Elsenbroich. Explanation in agent-based modelling: functions, causality or mechanisms? *Journal of Artificial Societies and Social Simulation*, 15(3):1, 2012.

[56] Brian Epstein. Agent-Based Modeling and the Fallacies of Individualism. In *Models, Simulations, and Representations*. Routledge, 2011.

[57] Joshua M Epstein. Agent-based computational models and generative social science. *Complexity*, 4(5):41–60, 1999.

[58] Joshua M Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling (Princeton Studies in Complexity)*. Princeton University Press, January 2007.

[59] Joshua M Epstein. Why model? *Journal of Artificial Societies and Social Simulation*, 11(4):12, 2008.

[60] Annalisa Fabretti. Calibration of an agent based model for financial markets. In Javier Pérez, Juan Corchado, María Moreno, Vicente Julián, Philippe Mathieu, Joaquin Canada-Bago, Alfonso Ortega, and Antonio Caballero, editors, *Highlights in Practical Applications of Agents and Multiagent Systems*, volume 89 of *Advances in Intelligent and Soft Computing*, pages 1–8. Springer Berlin / Heidelberg, 2011.

[61] Annalisa Fabretti. On the problem of calibrating an agent based model for financial markets. *Journal of Economic Interaction and Coordination*, June 2013.

[62] Nils Ferrand. *Modelling with and for Stakeholders*. GEMAS Studies in Social Analysis. Bardwell, 2007.

[63] DB Fogel. *Evolutionary computation: toward a new philosophy of machine intelligence*, volume 1. John Wiley & Sons, 2006.

[64] Floyd J Fowler. *Survey Research Methods*. Sage, 2013.

[65] Christopher Fox, Anany Levitin, and Thomas Redman. The notion of data and its quality dimensions. *Information Processing & Management*, 30(1):9–19, 1994.

[66] A Franklin, M Anderson, D Brock, S Coleman, J Downing, A Gruvander, J Lilly, J Neal, D Peterson, M Price, R Rice, L Smith, S Speirer, and D Toering. Can a theory-laden observation test the theory? *The British Journal for the Philosophy of Science*, 40(2):pp. 229–231, 1989.

[67] Mathias Franz and Charles L Nunn. Network-based diffusion analysis: a new method for detecting social learning. *Proceedings. Biological Sciences / The Royal Society*, 276(1663):1829–36, May 2009.

[68] Roman Frigg and Julian Reiss. The philosophy of simulation: hot new issues or same old stew? *Synthese*, 180(1):77, 2011.

[69] Rubén Fuentes-Fernández, Samer Hassan, Juan Pavón, José Galán, and Adolfo López-Paredes. Metamodels for role-driven agent-based modelling. *Computational & Mathematical Organization Theory*, 18(1):91–112, 2012.

[70] José Manuel Galán, Luis R Izquierdo, Segismundo S Izquierdo, José Ignacio Santos, Ricardo del Olmo, Adolfo López-Paredes, and Bruce Edmonds. Errors and artefacts in agent-based modelling. *Journal of Artificial Societies and Social Simulation*, 12(1):1, 2009.

[71] Rosanna Garcia, Paul Rummel, and John Hauser. Validating agent-based marketing models through conjoint analysis. *Journal of Business Research*, 60(8):848–857, 2007.

[72] Joés Garrido. Models and simulation. In *Object Oriented Simulation*, pages 1–16. Springer US, 2009.

[73] Nigel Gilbert. Varieties of emergence. In *Agent 2002 Conference: Social agents: Ecology, Exchange, and Evolution*, 2002.

[74] Nigel Gilbert. *Agent-Based Models*. Quantitative applications in the social sciences. Sage, Los Angeles, CA, 2008.

[75] Nigel Gilbert and Pietro Terna. How to build and use agent-based models in social science. *Mind & Society*, 1(1):57–72, 2000.

[76] Nigel Gilbert and Klaus Troitzsch. *Simulation for the Social Scientist*. Open University Press, 2005.

[77] Marco Gillies. Learning finite-state machine controllers from motion capture data. *Computational Intelligence and AI in Games, IEEE*, 1(1):63–72, 2009.

[78] Vincent Ginot and Herve Monod. *Exploring Models by Simulation*. GEMAS Studies in Social Analysis. Bardwell, 2007.

[79] Robert Goldfarb, Thomas C Leonard, and Steven M Suranovic. Are rival theories of smoking underdetermined? *Journal of Economic Methodology*, 8(2):229–251, 2001.

[80] John H Goldthorpe. *Causation, Statistics and Sociology*. Economic and Social Research Institute (ESRI), 1998.

[81] Diana F Gordon and Marie Desjardins. Evaluation and selection of biases in machine learning. *Machine Learning*, 17:1–17, 1995.

[82] Christina Goulding. Grounded theory, ethnography and phenomenology: A comparative analysis of three qualitative strategies for marketing research. *European Journal of Marketing*, 39(3/4):294–308, 2005.

[83] Ursula Grandcolas, Ruth Rettie, and Kira Marusenko. Web survey bias: sample or mode effect? *Journal of Marketing Management*, 19(5-6):541–561, 2003.

[84] William H Greene. *Econometric Analysis*. Prentice Hall, New Jersey, fifth edition, 2002.

[85] Till Grune-Yanoff. The explanatory potential of artificial societies. *Synthese*, 169(3):539–555, 2009.

[86] Paul Hague, Nick Hague, and Carol-Ann Morgan. *Market Research in Practice*. Kogan Page, 2004.

[87] David J Hand and William Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.

[88] D Wade Hands. *Reflection without Rules*. Cambridge University Press, 2001.

[89] Shaun Hargreaves-Heap and Yanis Varoufakis. *Game Theory: A Critical Introduction*. Routledge, 2 edition, April 2004.

[90] Samer Hassan and Javier Arroyo. Asking the Oracle: Introducing forecasting principles into agent-based modelling. *Journal of Artificial Societies and Social Simulation*, 16(3):13, 2013.

[91] Samer Hassan, Juan PavÃn, Luis Antunes, and Nigel Gilbert. Injecting Data into Agent-Based Simulation. In Hiroshi Deguchi, Shu-Heng Chen, Claudio Cioffi-Revilla, Nigel Gilbert, Hajime Kita, Takao Terano, Keiki Takadama, and Guillaume Deffuant, editors, *Simulating Interacting Agents and Social Phenomena*, volume 7 of *Springer Series on Agent Based Social Systems*, pages 177–191. Springer Japan, 2010.

[92] Brian Heath, Raymond Hill, and Frank Ciarallo. A survey of agent-based modeling practices (January 1998 to July 2008). *Journal of Artificial Societies and Social Simulation*, 12(4):9, 2009.

[93] David Hendry. *Econometrics: Alchemy or Science?* Blackwell Publishers, Oxford UK, 2000.

[94] David F Hendry and Hans-Martin Krolzig. Automatic model selection: a new instrument for social science. *Electoral Studies*, 23(3):525–544, 2004.

[95] Marko Hofmann. Ontologies in modeling and simulation: an epistemological perspective. In *Ontology, Epistemology, and Teleology for Modeling and Simulation*, pages 59–87. Springer, 2013.

[96] Kevin D Hoover. The methodology of econometrics. *New Palgrave Handbook of Econometrics*, 1(January), 2006.

[97] Robert Huckfeldt, Carol Kohfeld, and Thomas W Likens. *Dynamic Modelling - An Introduction (Quantitative Applications in the Social Sciences)*. SAGE Publications, 1982.

[98] Paul Humphreys and Cyrille Imbert. *Models, Simulations, and Representations*. Routledge studies in the philosophy of science. Routledge,, New York :, 2011.

[99] Peter Jackson. *Introduction to Expert Systems*. Addison Wesley, 3rd edition, December 1998.

[100] Wander Jager. The four P's in social simulation, a perspective on how marketing could benefit from the use of social simulation. *Journal of Business Research*, 60(8):868–875, August 2007.

[101] Marco Janssen. Evolution of cooperation when feedback to reputation scores is voluntary. *Journal of Artificial Societies and Social Simulation*, 9(1):17, 2006.

[102] Marco A Janssen. Understanding Artificial Anasazi. *Journal of Artificial Societies and Social Simulation*, 12(4):13, 2009.

[103] Xiongbing Jin and Roger White. An agent-based model of the influence of neighbourhood design on daily trip patterns. *Computers, Environment and Urban Systems*, 36(5):398–411, September 2012.

[104] Barbara E Kahn, Donald G Morrison, and Gordon P Wright. Aggregating individual purchases to the household level. *Marketing Science*, 5(3):pp. 260–268, 1986.

[105] Daniel Kahneman. Maps of bounded rationality: psychology for behavioral economics. *The American Economic Review*, 93(5):pp. 1449–1475, 2003.

[106] Andreas M. Kaplan and Michael Haenlein. Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. *Business Horizons*, 54(3):253–263, May 2011.

[107] Bonnie Kaplan and Dennis Duchon. Combining qualitative and quantitative methods in information systems research: a case study. *MIS Quarterly*, 12(4):pp. 571–586, 1988.

[108] Bonnie Kaplan and Joseph A Maxwell. Qualitative research methods for evaluating computer information systems. *Evaluating the Organizational Impact of Healthcare Information Systems*, pages 30–55, 2005.

[109] Peter Kennedy. *A Guide to Econometrics*, volume 2. MIT Press, 2008.

[110] Hugo A Keuzenkamp and Jan R Magnus. On tests and significance in econometrics. *Journal of Econometrics*, 67(1):5–24, 1995.

[111] Elmar Kiesling, Markus Günther, Christian Stummer, and Lea M. Wakolbinger. Agent-based simulation of innovation diffusion: a review. *Central European Journal of Operations Research*, 20(2):183–230, May 2011.

[112] Jan H Kietzmann, Bruno S Silvestre, Ian P McCarthy, and Leyland F Pitt. Unpacking the social media phenomenon: towards a research agenda. *Journal of Public Affairs*, 12(2):109–119, 2012.

[113] Jae-Woo Kim. A tag-based evolutionary prisoner's dilemma game on networks with different topologies. *Journal of Artificial Societies and Social Simulation*, 13(3):2, 2010.

[114] Gary King. Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721, 2011.

[115] Ben Klemens. *Modeling with Data: Tools and Techniques for Scientific Computing*. Princeton University Press, Princeton, N.J., 2008.

[116] David Knoke and Song Yang. *Social Network Analysis*, volume 154 of *Quantitative Applications in the Social Sciences*. Sage, Los Angeles, CA, 2nd ed edition, 2008.

[117] Philip Kotler. *Marketing Management*. Prentice Hall, millenium edition, 2000.

[118] Philip Kotler and Kevin L Keller. *Marketing Management*. Pearson/Prentice Hall, 2006.

[119] Thomas S Kuhn. *Objectivity, Value Judgment, and Theory Choice*, pages 320–339. Number 13. University of Chicago Press, 1977.

[120] John E Laird. Extending the Soar cognitive architecture. In *Frontiers in Artificial Intelligence and Applications*, pages 224–235, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[121] Pat Langley, John E Laird, and Seth Rogers. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160, June 2009.

[122] Lauri Lättilä, Per Hilletofth, and Bishan Lin. Hybrid simulation models — when, why, how? *Expert Systems with Applications*, 37(12):7969–7975, 2010.

[123] Michael R Lavington. A practical microsimulation model for consumer marketing. *Operational Research Quarterly (1970-1977)*, 21(1):pp. 25–45, 1970.

[124] David Lazer. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[125] Kang Hoon Lee, Myung Geol Choi, Qyoun Hong, and Jehee Lee. Group behavior from video: a data-driven approach to crowd simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '07, pages 109–118, Aire-la-Ville, Switzerland, Switzerland, 2007. Eurographics Association.

[126] Shuliang Li, Jim Zheng Li, Hong He, Philippa Ward, and Barry J. Davies. WebDigital: A Web-based hybrid intelligent knowledge automation system for developing digital marketing strategies. *Expert Systems with Applications*, 38(8):10606–10613, August 2011.

[127] Youwei Li, Bas Donkers, and Bertrand Melenberg. Econometric analysis of microscopic simulation models. *Quantitative Finance*, 10(10):1187–1201, December 2010.

[128] Henry Lieberman. Letizia: An agent that assists web browsing. *IJCAI (1)*, 1995.

[129] Linda Lilburne and Stefano Tarantola. Sensitivity analysis of spatial models. *International Journal of Geographical Information Science*, 23(2):151–168, February 2009.

[130] Gary L Lilien. Bridging the academic–practitioner divide in marketing decision models. *Journal of Marketing*, 75(4):196–210, June 2011.

[131] Gary L Lilien, Philip Kotler, and K. Sridhar Moorthy. *Marketing Models*. Prentice-Hall of India, 2003.

[132] Pierre Livet. *Towards an Epistemology of Multi-agent Simulation in Social Sciences*. GEMAS Studies in Social Analysis. Bardwell, 2007.

[133] Pierre Livet, Jean-Pierre Muller, Denis Phan, and Lena Sanders. Ontology, a mediator for agent-based modeling in social science. *Journal of Artificial Societies and Social Simulation*, 13(1):3, 2010.

[134] Michael Luck, Steve Munroe, Ronald Ashri, and Fabiola López y López. Trust and norms for interaction. In *SMC (2)*, pages 1944–1949, 2004.

[135] George F Luger. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Addison-Wesley Publishing Company, USA, 6th edition, 2008.

[136] Thomas Lux. Estimation of an agent-based model of investor sentiment formation in financial markets. *Journal of Economic Dynamics and Control*, 36(8):1284–1302, August 2012.

[137] Michael W Macy and Robert Willer. From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, 28:pp. 143–166, 2002.

[138] Franco Malerba, Richard Nelson, Luigi Orsenigo, and Sidney Winter. History-friendly models of industry evolution: the computer industry. *Industrial and Corporate Change*, 8(1):3–40, March 1999.

[139] Gianluca Manzo. Variables, mechanisms, and simulations : Can the three methods be synthesized? *Revue Francaise de Sociologie*, 48(-):35–71, 2007.

[140] Robert Ernest Marks. Validating simulation models: A general framework and four applied examples. *Computational Economics*, 30(3):265–290, October 2007.

[141] Stephen Marsland. *Machine Learning: An Algorithmic Introduction*. CRC Press, New Jersey, USA, 2009.

[142] Zbigniew Michalewicz. *Genetic algorithms+ data structures= evolution programs*. 1996.

[143] David Midgley, Robert Marks, and Dinesh Kunchamwar. Building and assurance of agent-based models: An example and challenge to the field. *Journal of Business Research*, 60(8):884–893, 2007.

[144] David F Midgley, Robert E Marks, and Lee G Cooper. Breeding competitive strategies. *Management Science*, 43(3):257–275, 1997.

[145] Marvin Minsky. Matter, mind and models. In *Proceedings of IFIP Congress*, pages 45–49, 1965.

[146] Tom M Mitchell. Mining our reality. *Science*, 326(5960):1644–1645, December 2009.

[147] Christopher Z Mooney. *Monte Carlo Simulation (Quantitative Applications in the Social Sciences)*. SAGE Publications, April 1997.

[148] Robert E. Morgan. Conceptual foundations of marketing and marketing theory. *Management Decision*, 34(10):19–26, 1996.

[149] Scott Moss. Social simulation models and reality: Three approaches. In Jaime S Sichman, Rosaria Conte, and Nigel Gilbert, editors, *Multi-Agent Systems and Agent-Based Simulation*, volume 1534 of *Lecture Notes in Computer Science*, pages 45–60. Springer Berlin / Heidelberg, 1998.

[150] Scott Moss and Bruce Edmonds. Sociology and simulation: Statistical and qualitative cross-validation. *American Journal of Sociology*, 110:1095–1131, 2005.

[151] Jean-Pierre Müller. Emergence of collective behaviour and problem solving. In Andrea Omicini, Paolo Petta, and Jeremy Pitt, editors, *Engineering Societies in the Agents World IV*, volume 3071 of *Lecture Notes in Computer Science*, page 519. Springer Berlin / Heidelberg, 2004.

[152] Jean-Pierre Müller. Mimosa: Using ontologies for modelling and simulation. In *GI Jahrestagung (1)*, pages 227–231, 2007.

[153] SA Neslin, Caroline Henderson, and John Quelch. Consumer promotions and the acceleration of product purchases. *Marketing Science*, 4(2):147–165, 1985.

[154] Michael J North, Charles M Macal, James St. Aubin, Prakash Thimmapuram, Mark Bragen, June Hahn, James Karr, Nancy Brigham, Mark E Lacy, and Delaine Hampton. Multiscale agent-based consumer market modeling. *Complexity*, 15(5):37–47, 2010.

[155] Martin A Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, December 2006.

[156] Martin A Nowak and Karl Sigmund. Evolution of indirect reciprocity. *Nature*, 437:1291–1298, October 2005.

[157] James J Odell, H Van Dyke Parunak, Mitch Fleischer, and Sven Brueckner. Modeling agents and their environment. In *Proceedings of the 3rd International Conference on Agent-oriented Software Engineering III*, AOSE'02, pages 16–31, Berlin, Heidelberg, 2003. Springer-Verlag.

[158] Scott Page. Aggregation in agent-based models of economics. *Knowledge Engineering Review*, 27(1):151–162, 2012.

[159] H. Van Dyke Parunak, Robert Savit, and Rick L Riolo. Agent-based modeling vs. equation-based modeling: A case study and users' guide. In *Multi-Agent Systems and Agent-Based Simulation*, pages 10–25, London, UK, 1998. Springer-Verlag.

[160] Renana Peres, Eitan Muller, and Vijay Mahajan. Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing*, 27(2):91–106, June 2010.

[161] Teodosio Perez-Amaral, Giampiero M Gallo, and Halbert White. A comparison of complementary automatic modeling methods: RETINA and PcGets. *Econometric Theory*, 21(01):262–277, 2005.

[162] Denis Phan and Frédéric Amblard, editors. *Agent-based Modelling and Simulation in the Social and Human Sciences*. GEMAS Studies in Social Analysis. Bardwell, 2007.

[163] Ivan Png and Dale Lehman. *Managerial Economics*. Blackwell Publishing, 2007.

[164] Philip M Podsakoff and Scott B MacKenzie. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5):879, 2003.

[165] Sarvapali D Ramchurn, Dong Huynh, and Nicholas R Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.

[166] William Rand and Roland T Rust. Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, 28(3):181–193, 2011.

[167] Anand Rao. AgentSpeak(L): BDI agents speak out in a logical computable language. In Walter de Velde and John Perram, editors, *Agents Breaking Away*, volume 1038 of *Lecture Notes in Computer Science*, pages 42–55. Springer Berlin / Heidelberg, 1996.

[168] Rick L Riolo, Michael D Cohen, and Robert Axelrod. Evolution of cooperation without reciprocity. *Nature*, 414:441–443, 2001.

[169] John H Roberts and James M Lattin. Development and testing of a model of consideration set composition. *Journal of Marketing Research*, 28(4):pp. 429–440, 1991.

[170] John H Roberts and Gary L Lilien. Explanatory and predictive models of consumer behavior. In J Eliashberg and G L Lilien, editors, *Marketing*, volume 5 of *Handbooks in Operations Research and Management Science*, pages 27–82. Elsevier, 1993.

[171] John H Roberts, Charles J Nelson, and Pamela D Morrison. A prelaunch diffusion model for evaluating market defense strategies. *Marketing Science*, 24(1):150–164, August 2005.

[172] Stuart J Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 3. edition, 2010.

[173] Andrea Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280–297, May 2002.

[174] Andrea Saltelli. Global sensitivity analysis: an introduction. *Proc. 4th International Conference on Sensitivity Analysis of Model Output (SAMO04)*, pages 27–43, 2004.

[175] Candelaria Sansores and Juan Pavón. Agent-based simulation replication: A model driven architecture approach. In Alexander Gelbukh, Álvaro de Albornoz, and Hugo Terashima-Marín, editors, *MICAI 2005: Advances in Artificial Intelligence*, volume 3789 of *Lecture Notes in Computer Science*, pages 244–253. Springer Berlin / Heidelberg, 2005.

[176] Thomas C Schelling. *Micromotives and Macrobehavior*. W. W. Norton, revised edition, 1978.

[177] Arndt Schwaiger and Bjrn Stahmer. SimMarket: Multiagent-based customer simulation and decision support for category management. In Michael Schillo, Matthias Klusch, Jrg Mller, and Huaglory Tianfield, editors, *Multiagent System Technologies*, volume 2831 of *Lecture Notes in Computer Science*, pages 1094–1096. Springer Berlin / Heidelberg, 2003.

[178] Sandip Sen and Stéphane Airiau. Emergence of norms through social learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1507–1512, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[179] Abhijit Sengupta and Stephen Glavin. Volatility in the consumer packaged goods industry a simulation based study. *Advances in Complex Systems*, 13(04):579–605, 2010.

[180] Abhijit Sengupta and Danica Vukadinović Greetham. Dynamics of brand competition: Effects of unobserved social networks. *Journal of Economic Dynamics and Control*, 34(12):2391–2406, 2010.

[181] Cosma Rohilla Shalizi. Methods and techniques of complex systems science: An overview. In Thomas S Deisboeck and J Yasha Kresh, editors, *Complex Systems science in Biomedicine*, Topics in Biomedical Engineering International Book Series, chapter 2, pages 33–114. Springer US, 2006.

[182] Xuhui Shao and Lexin Li. Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 258–264, New York, NY, USA, 2011. ACM.

[183] Robert Shaw and David Merrick. *Marketing Payback*. Prentice Hall, 2005.

[184] Jeff Shrager and Pat Langley. *Computational models of scientific discovery and theory formation*. Morgan Kaufmann Publishers, San Mateo, California, 1990.

[185] Michael Silberstein and John McGeever. The search for ontological emergence. *The Philosophical Quarterly*, 49(195):201–214, 1999.

[186] Munindar P Singh. Agent communication languages: Rethinking the principles. *Computer*, 31:40–47, 1998.

[187] Martin Spielauer. What is social science microsimulation? *Social Science Computer Review*, 29(1):9–20, 2011.

[188] Forrest Stonedahl and Uri Wilensky. Finding forms of flocking: Evolutionary search in abm parameter-spaces. *Multi-Agent-Based Simulation XI*, pages 61–75, 2011.

[189] Ryan Sullivan, Allan Timmermann, and Halbert White. Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics*, 105(1):249–286, 2001.

[190] Eric Sun, Itamar Rosenn, Cameron Marlow, and TM Lento. Gesundheit! Modeling contagion through Facebook news feed. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S Glance, Nicolas Nicolov, and Belle L Tseng, editors, *ICWSM*. The AAAI Press, 2009.

[191] Dicky Suryadi and Piotr J Gmytrasiewicz. Learning models of other agents using influence diagrams. *Courses and Lectures-International Centre for Mechanical Sciences*, pages 223–234, 1999.

[192] Charles S Taber and Richard J Timpone. *Computational Modeling (Quantitative Applications in the Social Sciences)*. Sage Publications, Inc, March 1996.

[193] Gerard J. Tellis. Modeling marketing mix. *Handbook of Marketing Research*, pages 506–522, 2006.

[194] Takao Terano. Exploring the vast parameter space of multi-agent based simulation. In *Proceedings of the 2006 International Conference on Multi-Agent-Based Simulation VII*, MABS'06, pages 1–14, Berlin, Heidelberg, 2007. Springer-Verlag.

[195] Paul Thagard and Ziva Kunda. Making sense of people: Coherence mechanisms, 1998.

[196] Paul Thagard and Karsten Verbeurgt. Coherence as constraint satisfaction. *Cognitive Science*, 22(1):1–24, January 1998.

[197] Nicholas S Thompson and Patrick Derr. Contra Epstein, good explanations predict. *Journal of Artificial Societies and Social Simulation*, 12(1):9, 2009.

[198] Ljupčo Todorovski and Sašo Džeroski. Automatic tuning of agent-based models using genetic algorithms. *Ecological Modelling*, 194(1-3):3–13, March 2006.

[199] Dionysios C Tsichritzis and Frederick H Lochovsky. *Data Models*. Prentice Hall Professional Technical Reference, 1982.

[200] John G Wacker. A definition of theory: research guidelines for different theory-building research methods in operations management. *Journal of Operations Management*, 16(4):361–385, 1998.

[201] Duncan J Watts and Peter S Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, December 2007.

[202] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[203] Danny Weyns, Andrea Omicini, and James Odell. Environment as a first class abstraction in multiagent systems. *Autonomous Agents and Multi-Agent Systems*, 14(1):5–30, 2007.

[204] Patrick Wiegand, Randy Pell, and Enric Comas. Simultaneous variable selection and outlier detection using a robust genetic algorithm. *Chemometrics and Intelligent Laboratory Systems*, 98(2):108–114, October 2009.

[205] Paul Windrum. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2):8, 2007.

[206] Michael Wooldridge. *An Introduction to Multiagent Systems*. Wiley, Chichester, UK, 2nd edition, 2009.

[207] Michael Wooldridge and NR Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

[208] Danny Wyatt. Collective modeling of human social behavior. *AAAI Spring Symposium: Human Behavior Modeling*, pages 86–91, 2009.

[209] Lu Yang and Nigel Gilbert. Getting away from numbers: Using qualitative observation for agent-based modeling. *Advances in Complex Systems*, 11(2):175–185, 2008.

[210] Bai Yu and Hong Fei. Modeling social cascade in the Flickr social network. In *Fuzzy Systems and Knowledge Discovery, 2009*, pages 566–570, Washington, DC, USA, August 2009.

[211] Gloria L Zeniga. An ontology of economic objects: An application of Carl Menger's ideas. *American Journal of Economics and Sociology*, 58(2):pp. 299–312, 1999.

[212] Tao Zhang and David Zhang. Agent-based simulation of consumer purchase decision-making and the decoy effect. *Journal of Business Research*, 60(8):912–922, August 2007.

[213] Peter Zubcsek and Miklos Sarvary. Advertising to a social network. *Quantitative Marketing and Economics*, 9(1):71–107, 2011.